DOCTORAL DISSERTATION

KISS ILONA

A GRAND TOUR

Investigations into the Self-compiled Tourism English Corpus and its Potentials in Teaching
the Vocabulary and Discourse of Tourism

2023

Doctoral Programme in English Applied Linguistics

University of Pécs

KISS ILONA

A GRAND TOUR

Investigations into the Self-compiled Tourism English Corpus and its Potentials in Teaching the Vocabulary and Discourse of Tourism

Supervisor: József Horváth, PhD, habil

Date of submission: 2023

Témavezetői nyilatkozat a dolgozat benyújtásához

Alulírott Dr Horváth József nyilatkozom, hogy Kiss Ilona doktorjelölt

Investigations into the Self-compiled Tourism English Corpus and its Potentials in Teaching the Vocabulary and Discourse of Tourism

című doktori értekezését megismertem, nyilvános vitára bocsátását támogatom.

Dátum: Pécs, 2023. augusztus 29.

....
témavezető aláírása

Table of contents

Acknowledgements

I am especially grateful to my supervisor PhD habil Horváth József for his profound academic guidance and incessant personal support throughout my PhD studies, and for his invaluable remarks and suggestions on the drafts of the dissertation. I am also indebted to him for drawing my attention to corpora and their application in language teaching.

I would also like to express my gratitude to all my teachers who have contributed to the completion of this dissertation with their motivating and encouraging support.

I would like to thank my family, especially my son, and all my friends for their encouragement throughout my doctoral studies.

Abstract

The aim of this dissertation is to extend research on the application of self-compiled corpora in teaching the special language of tourism. As previous studies focussing on tourism students' vocabulary development with corpus methods and techniques are based on tourism corpora that are publicly not available, I examined the potentials of developing a specialized corpus, and compiling word and collocation lists that can be utilized in teaching English for tourism. The main aim of this study is to present in what ways a relatively small self-built corpus of tourism texts can be exploited to enhance learners' lexicon, and to develop their collocational awareness. In order to provide basis for the investigations on the compilation of word and collocation lists, on the structural and functional types of lexical bundles, on the semantic preferences of collocations, and on the persuasive discursive features of tourism language the so-called Tourism English Corpus (TEC) was compiled. The corpus-based and corpus-driven analyses focussed on frequent lexical items and multiword units of the tourism texts and on their collocational behaviour. The lexico-grammatical patterns were examined at lexical, syntactic and discursive levels, and scrutinized in terms of their grammatical structures and discourse functions. The investigations of tourism lexis were based on the notions of word family, lexical bundles and collocations, and yielded thematic word lists of the main domains of tourism, as well as collocation lists of the most frequently occurring verbs, nouns and adjectives in the corpus. The analysis of text coverage revealed that the lexis of the TEC texts is not highly technical, the majority of the lexical items of the word lists fall in the categories that are considered sub-technical and general vocabulary. The corpus-based analysis of persuasive discursive techniques in travel articles shed light on several specific features of the marketing-oriented tourism discourse, and presented some ways how lexico-grammatical choices as inherent persuasive devices affect the prospective customers' perception and decision making. The findings of the studies offer a novel contribution to different fields of research; they provide a deeper insight into the lexico-grammatical patterns and discursive features of the language of tourism, and present several methodological aspects of the compilation of a specialized corpus, word- and collocation lists, and their implementation in teaching the specialized vocabulary and language of tourism. They also provide a clearer understanding of tourism discourse, serving a basis for material and course design.

List of abbreviations

AVL	Academic Vocabulary List
AWL	Academic World List
AI	Artificial Intelligence
BASE	British Academic Spoken English Corpus
BICS	Basic Interpersonal Communication Skills
BNC	British National Corpus
CALL	Computer Assisted Language Learning
CBA	Content-based Approach
CBI	Content-based Instruction
CBLL	Content-based Language Learning
CEFR	Common European Framework of Reference for Languages
CHILDES	Child Language Data Exchange System
CL	Corpus Linguistics
CLIL	Content and Language Integrated Learning
CLT	Compleat Lexical Tutor
COBUILD	Collins and Birmingham University International Language Database
COCA	Corpus of American English
DDL	Data-Driven Learning
EAOP	English for Academic and Occupational Purposes
EAP	English for Academic Purposes
EBE	English for Business and Economics
EGP	English for General Purposes
ELFA	English as a Lingua Franca in Academic Settings
EMI	English as a Medium of Instruction for academics
EOP	English for Occupational Purposes
ESP	English for Specific Purposes
ESS	English for Social Sciences

EST	English for Science and Technology
ICE	International Corpus of English
ICLE	International Corpus of Learner English
KWIC	Key Word in Context
LLC	London-Lund Corpus of Spoken English
LOB	Lancaster-Oslo/Bergen Corpus
LSP	Languages for Specific Purposes
MICASE	Michigan Corpus of Academic Spoken English
MICUSP	Michigan Corpus of Upper- level Student Papers
NA	Needs Analysis
NP	Noun Phrase
PA	Pedagogical Agent
POS	Part of Speech
PP	Prepositional Phrase
SEU	Survey of English Language
TEC	Tourism English Corpus
TESL	Teaching English as a Second Language
TEFL	Teaching English as a Foreign Language
TITC	Tourism Industry Text Corpus
T-TourEC	Translational Tourism English Corpus
VOICE	Vienna - Oxford International Corpus of English
VP	Verb Phrase

INTRODUCTION

I.   The aim and shape of the thesis

With many years of experience in primary education, more than two decades ago I obtained a teacher of English position in a secondary vocational school of commerce and catering where a new training of tourism specialists was launched that year. Excited by the challenge of teaching English to students committed to different professions of tourism industry, my initial enthusiasm had gradually been overtaken by the concern of what and how to teach so that my learners would meet the requirements of their final exam. Yet, in spite of the obstacles to be overcome, I was captivated by the novelty and intricacy of the task, and by the opportunity of teaching professional language for would-be tour guides, travel agents and hotel personnel – besides general English. Since my learners were at an intermediate or advanced level, some of them started their tourism studies after taking B2 level proficiency exam, course objectives were targeted towards the acquisition of tourism terminology and discourse related to the main issues of this special field. However, encompassing the scope of a particular topic with its lexis seemed to remain a continual troublesome endeavour.

The pedagogical exploitation of corpora, large systematic collections of spoken and written texts stored in electronic format, and the application of corpus methods and tools seemed to be a reasonable and manageable alternative in teaching the special vocabulary and discourse of tourism. Despite the advances in corpus linguistics and the increasing number of researchers and practitioners who apply corpora as useful tools in their everyday work, corpora and corpus methods play a less important role in Hungarian language pedagogy, especially in teaching English for specific purposes (ESP), than their obvious and recognized merits would suggest. Corpora not only provide us with large amount of natural language examples and highlight what lexical items and multiword units are characteristic and specific in that language, but also enable practitioners to promote the acquisition of specialised lexicon and the development of professional communicative skills.

Drawing on theory and research in ESP, vocabulary teaching and corpus linguistics, the thesis attempts to integrate recent findings of these domains, and presents the findings of corpus-based and corpus-driven studies grounded on the results of my preliminary investigations. The

fundamental aim of the dissertation is to demonstrate in what ways the exploitation of a relatively small inventory of tourism texts, applying corpus methods and techniques, can be utilised to enhance learners' lexicon, and to develop their collocational competence. As the studies overarch a decade of research, the compilation of a larger-scale tourism corpus, pedagogically applicable word lists and collocation lists, and the investigation of the persuasive promotional discursive functions of tourism language have been additional aims of the study. Taking into account the learners' professional needs, the subcorpora cover the major topics of tourism English that are generally included in tourism proficiency exams.

The thesis addresses four fundamental questions. The first one concerns the feasibility, viability and applicability of a self-built tourism corpus that provides basis for the compilation of topic-related word lists, collocation lists, as well as to the corpus-driven lexico-grammatical analyses of travel articles. The second inquires into whether a useful inventory of tourism vocabulary can be specified related to a particular domain of tourism, and what potentials these subject-specific word lists offer in vocabulary teaching. The third focuses on what lexico-grammatical persuasive discursive tools can be detected in the texts of tourism articles. The fourth question addresses what structural and functional types of lexical bundles can be identified, and whether a pedagogically useful listing of collocations of the most frequent nouns, verbs and adjectives can be compiled.

The dissertation consists of two parts and comprises six chapters. Three chapters in **Part I, chapter 1, 2, and 3** provide general theoretical framework to the exploratory research on tourism discourse that combines quantitative and qualitative approaches, and to the compilation of the tourism corpus, and the word- and collocation lists.

Chapter 1 introduces the development and achievements of ESP and content-integrated approaches to language teaching, describing the key concepts of ESP, highlights some aspects of needs analysis, material and syllabus design, and discusses the similarities and differences of ESP and English for general purposes (EGP).

The second chapter is concerned with the key aspects of word knowledge and vocabulary acquisition. Different approaches to vocabulary knowledge are presented including the receptive/productive continuum and the access to the mental lexicon. The concept of formulaic language is discussed, the notion of collocation is clarified by giving an account of recent studies, as well as the characteristics of ESP vocabulary teaching and learning are highlighted.

An overview of the developments and recent achievements in corpus-based research is provided in the third chapter, focusing on the application of corpus methods in language pedagogy and vocabulary instruction. The basic concepts of corpus linguistics are introduced, as well as the stages of corpus compilation and analyses are presented in order to set the basis of the studies on word- and collocation lists, and on the discursive features of tourism language. The first part concludes that the integration of corpus methods and techniques in ESP vocabulary teaching and learning can contribute considerably to the improvement of learners' lexicon.

Part II presents the processes and methods of corpus-, wordlist-, and collocation list compilation and analyses, as well as the findings of the study on the persuasive and promotional discursive features of tourism language. The tourism subcorpora, collocation and keyword lists provide a useable dataset for the exploration of lexical patterns and extended lexico-grammatical analyses to both learners and teachers involved in English of tourism. The possibilities of the construction of a tourism corpus and thematic word lists are explored in Chapter 4, Chapter 5 presents the findings of a corpus-driven lexico-grammatical analysis of the discursive features of tourism language. Chapter 6 sets the context of the research on lexical bundles and collocations, by assessing the findings of a preliminary collocation study. It presents the process of the compilation of tourism collocation lists, and discusses the findings of the investigations on collocations and recurrent lexical units, concerning their syntactic structures, functional categories and semantic preferences. Chapter 4 and Chapter 6 look into the pedagogical implications of word- and collocation lists.

In the concluding section, the main findings of the studies are outlined, their limitations and shortcomings are discussed, as well as suggestions are provided for further development and investigations. The conclusions and pedagogical implications of the studies indicate further research directions for ESP practitioners who are concerned with the application of corpora and corpus methods, corpus- and wordlist compilation and analysis, as well as with corpus-based vocabulary instruction.

II.   Theoretical framework

Learners' needs in English for specific purposes (ESP) can be best defined as being able to use the language and professional communicative skills in occupational settings. For this reason, the thesis follows the approach of ESP and content-based instruction (CBI) research and teaching that focus on learners' special needs and interest. Since the main objective of the thesis is to investigate the potentials of developing tourism students' vocabulary and professional communicative skills in secondary vocational training, the theoretical framework is an integration of theories and research findings in ESP with focus on teaching/learning special language, underlying theories and approaches of vocabulary teaching and learning, corpus linguistics approach and its methodologies, as well as research on tourism language and discourse.

The corpus-based and corpus-driven studies are underlain by my keen interest in tourism discourse, in the specific features of tourism language, especially in its lexico-grammatical patterns and collocations. Consequently, the corpus-based investigations on the characteristic features of this specific language have been carried out with the ultimate aim in mind; the findings can be utilized and implemented in everyday teaching practices.

CHAPTER ONE

APPROACHES AND CHARACTERISTICS OF TEACHING ENGLISH FOR SPECIFIC PURPOSES

1.1. Introduction

The demand for communication in professional settings triggered the need for teaching and learning languages for specific purposes (LSP) due to the international scientific and economic expansion in the second half of the 20th century. With the worldwide burgeon of English as lingua franca, from the 1970s there has been a growing demand on English for specific purposes, that is, LSP applied to English, to meet the communicative requirements in business, science, technology and education. Since ESP lays special emphasis on language learners' specific communicative needs in professional and academic context (Long, 2005, p. 26), needs analysis is a defining element of its practices (Hyland, 2006, p. 380). In Hutchinson and Water's (1987, p. 19) view, ESP 'is an approach to language teaching in which all decisions as to content and method are based on the learners' reason for learning'. Needs analysis (NA) assigns and supports the main objective of ESP, the enhancement of learners' communicative competence that involves the knowledge of language and its appropriate use (Hymes, 1972) to achieve fluency in academic or occupational settings. However, NA does not pertain only to specific language learning needs but also to the long-term overall development of language skills (Case, 2013).

To accommodate language learners' needs, subject matter content became the starting point and organising principle of specific language courses, giving rise to theme-based instruction in the 1990s. The integration of language and subject matter content has received increased interest in the past three decades (Ardeo, 2013; Echevarria, Vogt & Short, 2004; Genesee, 1994; Grabe & Stoller, 1997; Knight, 2012; Pica, 2002; Wesche, 1993). As a consequence, a new approach, content-based instruction, emerged in language education in the 1990s grounded in the LSP approach, bilingual immersion programmes, and communicative language teaching.

CBI, and within its framework content and language integrated learning (CLIL), refer to innovative methodological approaches of curriculum and syllabus design that focus on developing proficiency by teaching a subject matter with and through a foreign language (Euridyce European Unit, 2006, p. 7). In CLIL, also called the content-based approach (CBA) (Ardeo, 2013, p. 26), learners are exposed to stimulating professional content and complete their academic tasks in the target language, thus acquire the language in a challenging and informative way.

The ESP and CBI approaches and methodologies share the assumption that in learning professional and academic discourse 'content/knowledge is inseparable from linguistic expression' (Ardeo, 2013, p. 27). The greatest merit of these approaches lies in providing opportunities to learners to acquire the target language through the specific subjects they have interest in (Bracaj, 2014, p. 40). Since the tasks are cognitively engaging, the integration of language learning and content learning results in increased motivation and interest, in addition, learners are aware of the fact that with their professional foreign language skills they have greater opportunities for employment (Grabe & Stoller, 1997).

The last decades have seen significant developments in ESP research, influenced by a text/discourse analysis perspective (Fujita, 2009, 2010; Gandin, 2014; Gotti, 2003; 2006). Genre- and corpus-based studies (Alcantar, 2007; Boulton et.al, 2012; Kang & Yu, 2011) have been in the focus of the main areas of ESP research, and new notions such as identity and intercultural communication have appeared within the framework of ESP. Globalization, the internationalization of higher education, and the role of English as a lingua franca in worldwide communication has led to a growing interest in the acquisition of profession-related and academic communicative skills in English.

The following section will highlight the main characteristics of teaching and learning English as a special language, and within its realm the key aspects of teaching and learning English in tourism are presented. The chapter reviews the development, definitions, classification, the approaches, main concepts and theoretical influences of ESP. The key features of the ESP and content-language integrated (CBI) approaches are discussed in the light of general language teaching, touching upon the role of ESP practitioners, presenting content-based teaching methods, and providing some considerations of material and course design. The notion of professional communicative competence is clarified, and some practical learning strategies are presented that can be applied to enhance specific communicative skills.

1.2. The ESP approaches

1.2.1. What the ESP approach involves

ESP is generally seen as a distinctive field within English language teaching that has developed its own procedures and methodology, sometimes by refraining from the trends of general language teaching (Dudley-Evans, 2000, p. 3). However, the main distinctive feature of ESP is merely the thoroughly contemplated profound needs analysis; although it is not unique to ESP, needs analysis facilitates preparing learners to communicate effectively in their work or study (Long, 2005, p. 26). ESP has evolved as a branch of English language teaching in response to the call for cost-effective specialised language courses and more subject-specific tailor-made teaching materials, and has ever since been at the cutting-edge of theory development and implementation of innovative practices in applied linguistics (Hyland, 2006, p. 379). In addition to the findings in applied linguistics, ESP teaching draws on research of various disciplines; this interdisciplinary perspective is another distinguishing characteristic of ESP (Bojovic, 2006, p. 487).

Interdisciplinarity, the openness to other approaches in a range of fields, is an underlying element of ESP practices. Besides its closest connections to applied linguistics, in particular with discourse analysis, there are strong links between ESP and pragmatics, critical literacy, and sociocognitive theory that holds the stance that an individual's knowledge acquisition is directly related to observing others in social interactions. The endeavour to embrace and unite different perspectives helps to identify what ESP aims at, assisting ESP practitioners and learners to interpret how the real communicative world works, and how these communicative practices can be implemented in ESP classes (Hyland, 2006, p. 380).

Due to the advances of technology, corpus linguistics has enabled the classification of massive linguistic data of different texts to identify the specific features of language (Belcher et. al. 2011, pp. 2-3). Dudley-Evans (2000, p. 10) argues that an influential approach to the theory of the ESP is based on the analysis of texts; texts of different genres in different disciplines have different linguistic features and patterns of organisation. In his stance, Swales (1990) has had a tremendous impact on identifying the linguistic features of different genres with the means of genre analysis, especially on teaching academic writing (Genre analysis is discussed in 1. 2. 2). Bojovic (2006) shares Dudley-Evans' (2000) view that 'the theory of ESP could be outlined based on the specific nature of the texts that learners need knowledge of, and on the need-related nature of teaching', (2006, p. 487). As a result of research on discourse and genres, ESP has

given insight into the structures and meanings of specific texts, discourse in academic or workplace contexts and into the pedagogic practices with which these communicative behaviours can be developed (Hyland, 2006, p. 380).

To refer to the five underlying basic principles of the ESP approach Swales (1990) used the term 'enduring conceptions', namely, (1) *authenticity*, (2) *research-base*, (3) *language-text*, (4) *learners' need*, and (5) *learning/methodology*. Authenticity, one of the main considerations and the earliest concept in the development of ESP, refers to both the application of authentic texts and authentic tasks. Swales (1990) demonstrated that research and pedagogy are intertwined; research generally has pedagogical purposes, and pedagogical practices are directed by research in a specified context. He asserted that on the basis of language use texts belong to particular genres according to their communicative purposes, and, ESP courses based on the findings and implementations of genre analysis have to meet learners' educational and occupational communicative needs (Swales, 1990, p. 58). Extending Swales' (1990) concepts of ESP, context, the sixth enduring conception, was introduced by Johns (2015, p. 114), who emphasised that persistent research is needed into both learners' needs and the target situation.

1.2.2. Overview of the development of the ESP approach

The first definitions of ESP appear relatively late in specialist literature, notwithstanding its presumed outset in the 1960s. Halliday, McIntosh and Strevens (1964) were the first to put an emphasis on the necessity and relevance of ESP. In their landmark work, *The linguistic sciences and language teaching (1964)*, they focussed on how language works and can be described as a human activity in specific forms of that language, and introduced the term 'register' to distinguish varieties of language based on their use. As proponents of 'the function helps determine form' view they established the research base and initiated discussions on special language teaching (DeMarco, 2011). Hutchinson and Waters (1987, p. 10) consider material writers Ewer and Latorre (1969) prominent descriptive ESP pioneers whose research was targeted to the analysis of large specialized corpora to outline the borderland of different registers with statistics.

ESP grew out of several converging trends to meet the specific communicative needs and practices of different social groups (Hyland, 2006, p. 379). The evolvement and extension of English for specific purposes in the 1970s and 80s is attributed to three reasons by Hutchinson

and Waters (1987, pp. 5-7); (1) the demand for English in professional context, (2) the developments in linguistics and (3) the achievements in education psychology. They share the view that two historic periods, the post-war era and the oil-crisis in the 70s contributed to the emergence of ESP. The extensive international economic, scientific and technical development and the influence of Western knowledge on oil-rich countries yielded the development of ESP, to satisfy the needs on job-related and academic courses. As English became an international language it created a new generation of ESP learners who had definite aims in language learning (Salmani-Nodoushan, 2007, p. 69), therefore, to cater their needs time and money constraints brought about the demand for cost-effective courses with clearly defined goals (Hutchinson & Waters, 1987, p. 7).

With the emergence of the communicative approach, pioneers in applied linguistics (Brumfit, 1980; Hymes, 1972; Wilkins, 1976) began to focus on the ways language is used in real-world situations (Widdowson, 1978), opposed to theoretical generative linguistics that described language in terms of grammar (Chomsky, 1966; Bloomfield, 1983). This revolutionary period of linguistics in the late 60s and early 70s witnessed the transformation of traditional linguistics into register analysis that highlighted the difference between written and spoken language (Javid, 2013, p. 144). Prominent linguists, such as Halliday (1980; 1988) and Swales (1985) supported the stance that the special English needed by a particular group of learners can be encompassed by analysing the linguistic characteristics of that occupational or academic field. The idea was developed one stage further; if language varies in different situations, to determine linguistic characteristics in order to meet learners' needs in different context seems also possible (Hutchinson & Waters, 1987, p. 7).

The third reason of the emergence of ESP is attributed to the achievements of educational psychology. The view on learners' real communicative needs was supported by new developments in education psychology; in contrast with the traditional language teaching methods, the learner-centered approach paid more attention to the process of language learning. Students use different learning strategies, and they are motivated by different needs and interest that have a marked impact on the effectiveness of their learning (Hutchinson & Waters, 1987, p. 7). With the help of tailor-made materials and targeted course design ESP teachers are often better equipped, to ensure that learners acquire the specific language, content, and the expected communication skills.

Since its early beginnings in the 1960s there have been five main stages in the development of ESP; (1) *register analysis*, (2) *rhetorical discourse analysis*, (3) *target language use situation analysis*, (4) *skills-centered approach*, and (5) *learning-centered approach* (Salmani-Nodoushan, 2007, pp. 70-71). *Register analysis* was in practice mainly in the 1960s and early 70s related to the works of corpus linguists Halliday, McIntosh and Strevens (1964), Ewer and Latorre (1969), and Swales (1971). Register analysis aimed to identify the grammatical and lexical features of different scientific registers, in ESP the analysis focussed on the sentence level.

The evolvement of discourse analysis led to the second phase of ESP development known as *rhetorical discourse analysis* that was most succinctly expressed by Allen and Widdowson (1974), who claimed that learners' difficulties arouse from their unfamiliarity of English use rather than from the deficiency of their knowledge of English as a system. They argued that ESP courses meet students' needs if they develop awareness of how sentences are used in performing different communicative acts. Consequently, discourse analysis has become a valuable research tool and important teaching method in ESP to exploit relevant authentic texts in classroom, in order to increase learners' awareness of their linguistic, discoursal and pragmatic features.

The interest in communicative language teaching and syllabus design yielded the evolvement of the third phrase, *target language use situation analysis* that aimed to establish procedures in which situation analysis is closely related to the learners' reason for learning. The identified features during the process of needs analysis provided basis to syllabus design in ESP courses (Hutchinson & Waters, 1987, pp. 9-12). *Skills-centered approach*, the fourth stage of ESP development, intended to investigate the thinking processes that underlie language use, assuming that learners did not need to focus on the forms of language but on the 'underlying interpretative strategies that enabled them to cope with the surface forms' (Salmani-Nodoushan, 2007, pp. 70-71). The shortcomings of the previous phases contributed to the emergence of the *learning-centered approach* in the fifth phase that concerned with the issue of what it means to know a language. Proponents of the approach as Davis, (1993), Felder and Brent (1996) argued that the preceding phases concerned only with language use, however, a valid approach to ESP must focus on the process of language learning.

To satisfy the need for authentic texts and useful teaching materials different approaches of text analysis gain prominence in ESP language teaching; *register analysis*, *rhetorical analysis*,

functional/notional approach, and the current dominant approach, *genre analysis*. These analyses and their results inform ESP instructors how language is used in different contexts, how the learning objectives of an ESP course could be adjusted to learners' needs, and sequenced (Todey, 2018, p. 2). Initially, the analysis of the ESP genres was restricted to *register analysis* that was associated with the identification of key grammatical and lexical elements in scientific texts (Swales, 1971). *Rhetorical analysis* introduced the idea that grammatical features found in specific context differ from the rules of general grammar books that have an impact on grammatical choice. The *functional/notional approach* is associated with course books; it is an effective method of teaching naturally co-occurring words, collocations. The greatest value of these approaches is that they evolved from the idea that texts used in specialist environments have particular features that distinguish them from other texts. These texts-based perspectives led to the development of genre analysis that integrates the achievements of the preceding approaches.

The impact of corpus linguistics has resulted in several methodological innovations in the field of ESP since the 1990s (Bowker & Pearson, 2002; Boulton, 2012, 2016; Conrad, 1999; Coxhead, 2002, 2012, 2013, 2017; Flowerdew, 1993, 2004; Gavioli, 2005). Corpus-based studies on genre analysis, academic writing, learning purpose and language use, learners' lexical needs, and discourse analyses revealed that there is a growing interest in this relatively new approach, and corpus analysis has become one of the most frequently used research approaches in the ESP domain. In addition, Data Driven Learning (DDL) has been used as an effective method in ESP classes to discover patterns in concordance lines. (Further relations of corpus linguistics and ESP will be discussed in 3.2.2., 3.2.3 and 3.2.4.)

New directions in ESP research were appraised by Belcher, Johns and Paltridge (2011, p. 3); they welcomed the tendency that ESP researchers had begun to investigate learner identity as central to needs analysis (Belcher & Lukkarila, 2011). A newly emerging approach, critical ethnography, has been applied to specific settings (Starfield, 2011), and to reveal ESP researchers' self-awareness during a needs analysis process (Johns & Makalela, 2011). To synchronize different approaches Belcher *et al*. (2011) advocated that the tendency towards the integration of research, language teaching and learning should be aligned with the advancement of ESP theory.

1.2.3. Definitions and key notions of LSP and ESP

Language for specific purposes (LSP) is framed by Swales (1992, p. 300) as 'the area of inquiry and practice in the development of language programs for people who need a language to meet a predictable range of communicative needs'. In Hyland's (2011) view LSP teaching refers to a distinctive approach to language education based on the identification of the specific features of that language, and of discourse and teaching practices that recognise the particular communicative, subject-matter needs, and expertise of learners. Hutchinson and Waters (1987) consider LSP a 'learning-centered approach' that builds on learners' previous knowledge and skills, promotes active learning, takes into account affective factors such as motivation and interest, and uses different means of input for communicative purposes.

ESP, as a subdivision of LSP, has been defined in various ways. However, as Antony (1997) pointed out, in spite of the various definitions there were considerable debates on its exact meaning in the 1970s and 80s. As Hutchinson and Waters (1987) pointed out 'ESP is *not* a matter of teaching specialised varieties of English', ESP must be seen as an approach, not as a product… It is an approach to language learning, which is based on the learner need (p. 19)'. They maintained that ESP can be best defined by saying what ESP is not, rather than by specifying what it is. Yet, a fairly comprehensive definition was provided by Strevens (1988, pp. 1-2), who distinguished four absolute characteristics and two variable characteristics to encompass the scope of ESP. In terms of absolute characteristics ESP is English language teaching which is:

> ➢ designed to meet specific needs of the learner,
> ➢ related in content to particular disciplines, occupations and activities,
> ➢ centered on the language appropriate to those activities in syntax, lexis, semantics
> ➢ in contrast with general English.

In terms of variable characteristics ESP may be, but not necessarily:

> ➢ restricted to the language skills to be learned (e. g., only speaking)
> ➢ not taught according to any pre-ordained methodology (pp. 1-2)

Strevens's definition identifies ESP in contrast to general English emphasising its relation to disciplines, occupations and activities, and designates the grounds of ESP courses. Dudley-

Evans (1997, p. 298) offered a modified definition based on that of Strevens, he improved it by removing the absolute characteristic that ESP is 'in contrast with General English'. Dudley-Evans and St. John (1998, pp. 4-5) added an absolute and five variable characteristics to those of Strevens, clarifying and extending his definition. According to this revised version, ESP makes use of the underlying methodology and activities of the discipline it serves, it may use a different methodology, and it is likely to be designed for adult learners either at a tertiary level institution or in a professional work situation. ESP could be taught and learnt, however, at secondary school level as well. As it is designed for intermediate or advanced students, most ESP courses assume some basic knowledge of the language system, nevertheless, Dudley-Evans warns (1997, pp. 9-11) that this definition is not necessarily related to a specific age group, or professional area.

The definitions of ESP highlight various characteristics of the approach, and entail the guiding principle of ESP that language teaching should promote the accomplishment of learners' specific needs and purposes. At its early stages, Mackay and Mountford (1978, p. 2) regarded ESP as teaching of English for 'utilitarian purposes'. Harding (2007) epitomized ESP in a practical working definition; it is 'the language of getting things done', claiming that in 'all definitions of ESP two elements are axiomatic: the sense of purpose and the sense of vocation' (p. 6). Johns (2015, pp. 114-115) accepts Dudley-Evans's and Hutchinson and Waters's characterizations, yet, she considers 'all good teaching ESP' emphasizing that 'most ESP teaching has been directed towards adult and near adult professionals and academics where it is the most effective'.

The integrative nature of ESP is emphasised in Robinson's (1991) summary: 'ESP is … an enterprise involving education, training and practice, and drawing upon three major realms of knowledge: language, pedagogy, and the students', participants', specialist areas of interest' (p. 1). She (1991) accentuated that ESP is goal-directed, thus, ESP courses develop from needs analysis and from a number of other learner and course-related characteristics that give reasonable explanations of the constraints in a limited period of time. Orr (2002) also highlights the aspect of time limit in ESP courses: ESP… is designed to meet the specific learning needs of a specific learner or group within a specific time frame', and views ESP as 'a branch of language education that studies and teaches subsets of English to assist learners in successfully carrying out specific tasks for specific purposes' (p. 207). Nevertheless, the various interpretations of ESP might depend on different viewpoints and geographical areas; as

Robinson (1991 p. 1) highlighted 'what is specific and appropriate in one part of the globe it may well not be elsewhere'.

1.2.4. ESP vs. EGP approach

As ESP and EGP (English for general purposes) overlap, one of the earliest challenges that ESP had to face was how to establish its distinctive domain in language teaching. EGP focuses on using English in general communicative situations, while ESP builds on EGP in preparing learners to attain specific professional or academic purposes (Salmani-Nodoushan, 2007, p. 69). Hutchinson and Waters (1987, pp. 53-54) claim that not the existence of learners' specific needs distinguishes ESP from EGP, but rather the awareness of that need and the target situation. For EGP teachers finding effective teaching materials for a particular class is a key concern that is also relevant in ESP, additionally, ESP teachers need to consider another factor, the acquisition of subject-specific knowledge (Day & Krzanowsky, 2011, p. 8).

To answer the question of what the fundamental difference is between the ESP and EGP teaching approach, Hutchinson and Waters (1987, p. 53) gave a simple explanation; 'in theory nothing, in practice a great deal'. Carver (1983) argued that all uses of English are specific in their nature, thus all teaching of English as a foreign language should be considered as language teaching to specific purposes. Hadley (2006, p. 3) asserted that the 'the key to ESP is the focus on the 'S' for specific'. The word *specific* might refer to both special language and specific needs and aims, thus, ESP can be distinguished from EGP by its concern to language related to disciplines and its practices. However, Gatehouse (2001) considers special language and specialized aims two entirely different notions; she argues that the word 'special' should focus on the purpose for which learners want to attend a course, not on the specific terminology or jargon that they learn.

In Belcher, Johns, and Paltridge's (2011, p. 1) stance, research on specific languages, discourses, contexts of use, students' academic and professional needs, and the subject-related pedagogical practices distinguish ESP from other branches of applied linguistics and approaches to language teaching. Basturkmen (2006) contrasts EGP with ESP setting different objectives:

> Whereas General English Language teaching tends to set out from point A toward an often pretty indeterminate destination, setting sail through largely uncharted waters, ESP aims to speed learners through to a known destination (p. 9).

Basturkmen supports the claim (2006, p. 9) that ESP is essentially a practical endeavor as the aim is to achieve the specific goals in the most time- and energy-saving manner. She emphasizes that although ESP is practical in its orientation, it is based on the same ideas about the nature of language and learning as any other approaches to language teaching.

Long (2005, p. 19) argues that general language courses teach vocabulary and skills that some learners do not need, while omit lexis and genres that they do. He proposes that instead of a 'one-size-fits-all approach' every language course should be viewed as involving specific purposes. For learners, however, language learning sometimes seems like 'language with no purpose', or 'language for other people's purposes' (p. 26). Addressing students' own specific language learning needs distinguishes ESP from other approaches to English language teaching; since teachers are aware of the importance of needs analysis, course objectives are the guiding principles of material development at all stages (Long, 2005, p. 26).

Drawing on Hutchinson and Water's (1987) definition Anthony (2018, p. 44) has specified the central components of ESP as 'four pillars', namely, *needs analysis*, *learning objectives*, *teaching materials and methods*, and the *evaluation of learners, instructors, courses and programs* in different ESP contexts. He emphasizes that different viewpoints such as diagnostic and analytic can be used to conduct needs analysis, and learning objectives can be appropriately sequenced in a course relying on the results of register and genre analyses, as they provide information how language is used in different contexts. Anthony (2018) advises the methodical evaluation of published materials and the creation of custom-made materials, and the use of technology and corpus tools. In his view (1997) needs analysis, course design, and material production demonstrate the greatest influence that the ESP approach has had on English teaching in general, and this way the dividing line between the EGP and ESP approaches blurs indeed.

1.2.5. Classification of ESP

Carver (1983) identifies three types of ESP: (1) *English as a restricted language*, (2) *English for Academic and Occupational Purposes (EAOP)*, and (3) *English with different topics*.

Restricted language refers to the language used in a vocational environment, for instance the language of air traffic controllers or waiters, this language would not allow effective communication outside this specific context (Gatehouse, 2001). The second type of ESP in Carver's (1983) division, *English for Academic and Occupational Purposes* occurs as different subcategories in Hutchinson and Waters's (1987, p. 17) tree of English language teaching.

Rooted on learning and communication Hutchinson and Waters (1987, p. 17) divided English as a subject-specific language into three branches:

➢ English for Science and Technology (EST)
➢ English for Business and Economics (EBE)
➢ English for Social Sciences (ESS)

Each of these subject areas is broken down into two further categories, *English for Academic Purposes (EAP)*, and *English for Occupational Purposes* (EOP). English for Economics is an example for EAP, whereas English for Secretaries is a subset of EOP. However, as Hutchinson and Waters (1987, p. 16) remark there is no clear-cut distinction between EAP and EOP, as people can work and study simultaneously. As Anthony remarks (2018) EAP and EOP exist on a 'continuum of needs that are weighted more strongly in one area or setting than another (p. 15)', and this way it is difficult to divide them into sub-categories. In addition, ESP is closely connected to the larger domains of Teaching English as a second language (TESL) and Teaching English as a foreign language (TEFL).

In the third type of ESP, English with different topics, the emphasis is laid on the topic not on the purpose. This topic-centered type concerns with the anticipated future needs of learners such as attending conferences, or writing research reports. Gatehouse (2001) does not regard English with different topics a separate type of ESP, but an integral part of ESP courses and programs with focus on situational language. (For more detailed typology of the various branches of ESP see Dudley Evans & St. John, 1998; Hutchinson & Waters, 1987; Mackay & Mountford, 1978).

English for tourism can be defined from the perspectives of *English for general Business Purposes vs. English for Specific Business Purposes*. Furthermore, English language teaching can be considered as a continuum with general English at one end and ESP at the other end (Dudley-Evans & St. John, 1998, pp. 8-9). Tourism English, a type of Business English, is closer to general English in most learning situations than other types of ESP; yet, it requires the

use of 'International English', i.e., the English used as a means of communication in business settings. English for tourism can be regarded as a type of either English for General or Specific Business Purposes depending on the teaching/learning situation; the former is characteristic at the very early stages of learners' career, the latter at the specialist level where the emphasis falls on Business English for Academic Purposes (Moreno, 2011, p. 140). English in tourism needs to fulfil different aims to meet the specific needs of tourism students and specialist; it can be further divided to sub-fields such as 'English for hotel industry', 'English for tourist management' or 'English for tourist guides'.

1.3. The learner-centered approach

1.3.1. ESP learners' competences

In the past decades there has been a shift in foreign language teaching from the attainment of objectives concerning grammatical and lexical knowledge, to the development of competencies. The core competencies to be acquired tend to fall into four categories: *generic*, *inter-cultural*, *linguistic* and *discourse competences* (Maican, 2014, p. 276). *Generic competences* including problem-solving, listening and persuasion skills have a lot in common with 'soft skills' or 'people skills' that give added value to students' or employees' performance as they are transferable from one field to the other. Generic competences are generally taught inductively in foreign language classes that are connected to appropriate language use; it follows that tourism students' generic competences include right attitude, tone, intonation, and professional communication skills. *Intercultural competence* refers to the values, behaviour and attitudes that help people communicate successfully with others from different cultures. *Linguistic competence* entails the acquisition of the basic elements of a language as phonology, grammar and vocabulary, whereas *discourse competence* involves the ability to produce appropriate language in different contexts (Maican, 2014, p. 276).

Learners taking an ESP course are generally motivated by specific interest, having acquired some subject matter knowledge and learning techniques that can enhance their performance in a particular field. Gatehouse (2001) posited that successful communication in an academic or occupational setting is possible only if students are able to communicate in everyday informal talk regardless of the context; their basic interpersonal communication skills (BICS) (Cummins, 1981) enable them to cope with the special content and vocabulary. In addition, they need to

learn the special lexis of the target language of a specific occupation or discipline, and acquire a set of academic or occupational skills such as conducting research or business correspondence.

In the time of globalization, developing learners' cultural awareness in ESP courses is a key to the success of effective professional communication; teachers of English have to pay special attention to the enhancement of learners' socio-pragmatic competence that entails how to apply interpersonal communicative strategies effectively, as cultural differences may have a significant impact on the ways of communication (Wisniewska, 2012). Sociolinguistic competence as an element of communicative competence entails not only lexical and grammatical knowledge, but also knowing how to use language appropriately in an academic or professional setting, and what words to use to express a specific attitude (Canale & Swain, 1980). As part of social competences, intercultural competence, 'the ability to behave and communicate effectively and appropriately in intercultural situations' (Deardorff, 2006, p. 196), is of paramount importance in the tourism sector that needs 'creative, knowledgeable specialists with a good command of several foreign languages, who are able to make decisions and work observing traditions of different cultures' (Luka, 2004, p. 2). At the same time, intercultural competence involves knowing the cultural values of students' own nation and country.

1.3.2. Teaching and learning ESP

Today, in European language teaching communities LSP programmes tend to adopt the practice that language teaching is not restricted to teaching words, lexical phrases, and grammatical structures in separation from the real context of use; instead, they foster the exploration of language that carries disciplinary and professional values. The language competences required by disciplines or occupations can be acquired through the understanding of the ways language forms and strategies work, represent and construct knowledge in a particular area (Hyland, 2011). To achieve the communicative and subject-specific goals of ESP courses, teachers need to use various kinds of profession-related materials that arouse students' motivation and interest to maximise the potentials of skill development in the course (Davies, 2003).

Tendencies in teaching ESP have also generated new challenges that teachers, who are expected to sustain learners' motivation and interest in class, have to overcome. As Belcher (2004) remarked: 'ESP pedagogy places heavy demands on its practitioners to collect empirical needs-assessment data, to create or adapt materials to meet specific needs identified, and to cope with

often unfamiliar subject matter and even language use… (p. 166)'. In the past decades, internationally more emphasis has been laid on material development, bridging the gap between course books and more relevant materials for a special group (Tomlinson, 2010; Wisniewska, 2012). In order to collect relevant materials students should be encouraged to participate in the analysis of their own specific needs (Basturkmen, 2010, p. 144). To augment teachers' expertise and skills, Holmes and co-authors (2023, p. 631) proposed collaborative teaching and learning with an Artificial Intelligence (AI) assistant that could automatically provide resources for classroom activities, including texts, images, videos or links, and monitor the students' performance in classroom activities and throughout the course as well. Although the constant monitoring of students' performance and achievement raises important ethical questions, it is conceivable that it will replace stop-and-test examinations.

To emphasize the multi-faceted role of the ESP teacher, the term 'practitioner' is proposed by Dudley-Evans and St John (1998, pp. 13-17). They identify five key roles for the ESP practitioner; *teacher*, *course designer and materials provider*, *collaborator*, *researcher* and *evaluator*. ESP teachers need to be flexible, take interest in the disciplines their teaching is involved in, design the course syllabus, and collect and assess teaching materials. ESP practitioners as researchers need to carry out needs analyses and be aware of the genres of the subject matter, and, as evaluators test not only the students' performance but assess the course and the teaching materials as well. However, as Bojovic (2006, p. 493) pointed out ESP teachers are not subject specialists, since learners generally have wider knowledge in the subject matter, they need to provide the framework, the necessary tools and materials to language skills development.

Yet, another problem, the training of ESP teachers and their willingness to teach ESP need to be solved in the future (Huang, 2010). According to Bojovic (2006, p. 493) ESP teachers often feel isolated from professionals and have difficulty in exchanging information in the field. Antony proposed (2009) cooperation with subject specialists as they can provide real world ESP experiences and can help learners with their precise, logical way of thinking on the subject. In order to address the needs of ESP teachers Hüttner *et al.* (2009) worked out the model of 'mediated corpus-based analysis for ESP teaching', integrating the theory and practice of applied linguistics, genre analysis, and corpus linguistics. This innovative approach enables pre-service teachers to act as English language teaching professionals, since it allows to develop the necessary competences, to familiarise themselves with unknown ESP genres, and operationalise them for teaching purposes.

Tourism is not only a fashionable professional area but it is the most popular topic with ESP practitioners according to a survey conducted in 2009 with Asian ESP college teachers; 23.46% of the practitioners asserted that they would teach a tourism course, whereas no one favoured English of natural science (Huang, 2010). However, working out the concept of courses, formulating objectives, and optimizing the process of language learning and the development of competencies place, high demands on language teachers. Simion (2012, p. 153) proposed not to teach English in tourism as a subject separated from students' real word, but integrating their experiences into the subject matter area they have interest in.

1.4. The content-based approach

1.4.1. Content-based instruction

The educational policy of the European Union, the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001), defines the main objectives and provides theoretical and practical guidelines to language teaching. One of the main educational aims of the CEFR is to enhance foreign language proficiency by direct exposure to authentic language use; teaching a subject matter through a foreign language is an ideal means of providing genuine use of that language. The *Action Plan* (2004-2006) of the Commission of European Communities fosters the participation in content-based instruction (CBI), namely, content language integrated learning (CLIL) courses that employ L2 as a medium of instruction, this way, contribute to the development of L2 vocabulary knowledge (Council of Europe, 2001, p. 2).

CLIL can be an effective approach and have positive outcomes on learners' language development at all stages of instruction. Stoller observed (2004) that students attending CLIL courses complete the courses with considerable content knowledge gain and improved language skills. The record of students' advancement, frequent comprehension checks, straightforward assignments and advancement procedures help master the content material and L2 mainly through incidental learning (Master, 1997).

Content-based instruction (CBI), as an organising principle of syllabus design, is often employed worldwide in theme-based courses of specific topic areas (Knight et. al., 2010, p. 4). The CBI approach, also called content-based language learning (CBLL), aims to develop 'use-oriented second and foreign language skills' and can be 'distinguished by the concurrent learning of a specific content and related language use skills' (Wesche, 1993, p. 81). In Krahnke's (1987, p. 65) definition CBI is 'the teaching of content or information in the language

being learned with little or no direct or explicit effort to teach the language itself separately from the content being taught'. The content is not necessarily academic or occupational; it can include any topic or theme that is important to learners or meet their interest (Genesee, 1994, p. 3). When information is presented in coherent meaningful context it leads to deeper processing and results in more effective learning that promotes retention; it is more likely that learners connect new pieces of information to related previous knowledge, thus content enhances language learning as well (Anderson, 1990). As content-language integrated learning activates learners' prior knowledge, they incorporate new words into their lexicon due to the formation of associative links around well-established words in their existing vocabulary.

As CBI and English as a medium of instruction for academics (EMI) integrate the content of the subject matter with the acquisition of the target language and communicative skills (Tarnopolsky, 2012, p. 73), they use 'language as a medium for learning content and content as a resource for learning and improving language' (Stoller, 2002, p. 109). Richards accentuated (2006, p. 28) that 'content provides a coherent framework that can be used to link and develop all the language skills', learners need to learn the content while their cognitive academic language proficiency develops (Cummins, 1981).

Since content learning is structured around academic and professional themes, high levels of competence can be achieved in CBI courses when learners gain information on topics of their needs and interest, and language skills are developed in parallel with mastering the main issues of a specific language (Larsen-Freeman, 2000, p. 142). It is easier to remember coherently presented thematically-organised information; when facts and skills are taught in isolation more practice and rehearsal are needed to internalize them in long-term memory (Singer, 1990). As CBI combines subject matter and language learning it can be highly motivating; students are able to apply what they learn in language classes to their professional fields of study, conversely, their performance in their subject matters can promote language learning (Fiorito, 2005).

1.4.2. ESP vs. content-based instruction

As opposed to ESP, CBI focuses on the content not on teaching the professional language, and thus CBI is not the evolution of ESP, yet, they both aim at achieving proficiency in an L2 and fluency in academic or occupational context (Ardeo, 2013. P. 29). A key difference between these two approaches is that in CBI content-learning objectives are equally or more important than language learning objectives, whereas in ESP language-learning objectives are

predominant. In Robinson's view (1991, pp. 2-4) in an ESP class language is a 'service' rather than a 'subject' for its own sake as language learning is integrated into a subject matter area. However, Pica (2002, pp. 15,39) warned that a strong focus on content and form can deprive learners of improving their proficiency, as language learning may be incidental, and errors are less often corrected.

In ESP, language is both the content and a means of learning that content, while in CBI 'scaffolding' techniques such as *verbal*, *procedural* and *instructional scaffolding* (Echevarria et.al., 2004, pp. 86-87) are advised to make the content more comprehensible and manageable (Ardeo, 2013. P. 29). Verbal scaffolding techniques such as paraphrasing, the use of cognates, synonyms and familiar lexical chunks promote language development in contextualized learning. Procedural scaffolding, the use of grouping techniques and activity frames, and instructional scaffolding that includes the use of materials and visual aids support content and language learning at the same time (Echevarria et.al, 2004, pp. 86-87). Yet, CBI and ESP share a number of features; the use of the content of different subjects, the development of academic and professional communication skills, the acquisition of special terminology, and the application of communicative language teaching methodology (Ardeo, 2013, p. 29).

To highlight potential pitfalls of subject-related language instruction, Macia and Bares (2015) investigated the role of CLIL and the implications derived from ESP programmes at university level. Findings showed that ESP course designers need to tailor courses to students' discipline-related needs to make them more relevant. They proposed collaboration with content lecturers to develop graduates' proficiency; it can take place either through the integration of language to content courses, or through the integration of content in ESP courses to make the content more relevant, and to meet students' academic and professional communicative needs.

1.4.3   Content-based models and teaching methods

The language learning objectives of a content-based program set the grounds for the foreign language and the content area curricula. Content-based syllabi hinge on the analytic approach to syllabus design; they offer a lot of exposure to the language, stimulating students to acquire the target language through using it. The benefits of content-based instruction are reflected in CLIL syllabi; the thematically organised materials are easier to remember, the information is presented coherently and meaningfully, and the topics develop as learners' knowledge progresses to complete more complex tasks (Jalilzadeh & Tahmasebi, 2014, pp. 225-230).

Different program models aim to eliminate the drawbacks of content-based language teaching to develop learners' proficiency within the framework of content-based syllabi. Besides the great values of CBI such as creating motivating and interesting language learning environment and developing communicative, cognitive, and collaborating skills, the implicit language instruction may give students the impression that they are not actually learning a language (Rhalmi, 2009). To adapt content-based instruction to learners' actual needs the most widely used approaches include (1) *sheltered model* (2) *theme-based instruction* (3) *adjunct model*, (4) *language across the curriculum*, and (5) *cognitive academic language learning approach* (Crandall, 1994; Davies, 2003).

Sheltered teaching involves two types of instruction: *sheltered subject-matter teaching* and *sheltered instruction*. In sheltered subject-matter teaching the language of texts, tasks and methods familiar to teachers are adapted to students' different proficiency levels, and the main objective of this type of content course is the mastery of content through incidental language learning. This language-sensitive content instruction is given by a content teacher or a language teacher with special expertise in the academic or professional area. When applying the sheltered instruction approach, the content curriculum is adjusted to learners' limited proficiency in an L2 (Crandall, 1994). (Investigations by Kiss and Horváth (2015) on the potential application of sheltered instruction in teaching tourism collocations will be presented in 6.6.3.).

In *theme-based* programmes the aim is to develop general academic language skills through relevant and engaging content (Crandall, 1994). A theme-based course can be taught either by a language teacher or in collaboration with a subject specialist (Davies, 2003). The *adjunct model* integrates a specific language learning course with a content course, in which both native speakers and second language learners are enrolled. The course is team-taught by a content teacher and a language teacher, thus, this model requires purposeful collaboration. The aim of the adjunct model is both the mastery of the content and language, the introduction to professional discourse, as well as developing transferable skills (Master, 1997).

Language across the curriculum and *cognitive academic language learning approach* are content-centered approaches that integrate language instruction into other curricular offerings (Crandall, 1994). The *language across the curriculum* method may include the development of integrated curricula and a kind of joint teaching of content and language instructors. The *cognitive academic language learning approach* is a combination of language, content, and

learning strategy instruction that forms a transitional approach to CBI for students at intermediate or advanced level of proficiency (Crandall, 1994).

1.5. Teaching and learning English in tourism

Due to the globalization and the Internet, the opportunities of travel and international communication have greatly enhanced, which resulted in the emergence of English in tourism. Tourism English, as a medium of tourism communication, has specific features and genres that tourism professionals need to acquire. The language of tourism attempts to seduce a great many of people to become tourists, influencing their attitude and behaviour. In turn, through communication, tourists and tourism professionals contribute to the formation and development of tourism language.

The demand for labour in the international tourism industry often results in the employment of not only national but foreign workforce who has to deal with visitors of different nations (Singh, 2008, p. 15), which, consequently, has increased the need for specific language skills of English in tourism. To satisfy the job-related communicative language needs of tourism students and employees there has been a growing demand for specific courses, resources, and research on tourism language and English in tourism. In teaching English for tourism language is not taught as a subject separated from the communicative needs of learners' real word, but integrated to a subject matter that brings their experiences and aspirations related to tourism to the classrooms (Simion, 2012, p.153).

Tourism students and specialists need to have a fairly good command of English and possess a high level of communication skills as tourism is a global and multilingual industry in which English has a dominant position (Francesconi, 2014, p. 10). To succeed in tourism careers tourism students should reach B2-C1 level according to the *Common European Framework for Languages* (CERF, 2001), as fluency in English is part of tourism occupational standards and an asset for employment and career advancement, as well as a basic requirement for achieving customers' satisfaction (Maican, 2014, p. 275). Accordingly, English in tourism focuses on the enhancement of professional communicative competence including the acquisition of the necessary technical vocabulary, and developing the awareness of typical genres related to the different fields of tourism. However, as Laborda (2009) observed, tourism students often lack a realistic perception of professional life and basic skills including communicative and negotiating skills.

Communicative competence is made up of sub-competences including linguistic competence that incorporates grammatical, lexical competence, strategic, sociolinguistic, and pragmatic competences that encompass contextual lexis (Luka, 2007, p. 17). Communicative competence involves *situative language use*, *genre competence*, knowing the specific genres of a professional field, *discourse competence*, and *register competence* that entails the knowledge of language elements used in professional situations (Rébék-Nagy, 2013, p. 6).

Grounded in language learning, students and specialists have to acquire ESP competence which is a combination of gained experience, attitude and abilities that allow them communicate in professional work, think strategically, implement their knowledge and work creatively (Luka, 2007, p. 16). The essential ESP competencies for tourism students embrace core competencies such as *cognitive competence* (practical and theoretical knowledge of the field), *personal competence* including communicative, intercultural and social competence, and *technologically-professional competence* that involves creative and constructive problem solving and cooperation. These competences can interact and be interdependent, based and developing on students' experiences. Tourism students' ESP competence develops in the teaching/learning process based on the collaboration and mutual understanding of teachers and group mates, and with the students themselves constructing their knowledge, since teaching methods and materials are adjusted to their needs (Luka, 2007, p. 9).

In contrast to teaching general English, teaching tourism English, within the realm of both ESP and CBI approaches, concentrates more on language use in special contexts than on teaching grammar and linguistic structures. Whereas in general language teaching the development of all four language skills is equally stressed, in tourism English the professional area determines which language skills are most needful for the students; the development of reading and writing skills are emphasised for the students in tourism administration, while the improvement of speaking skills is crucial for would-be tour guides and travel executives.

In Hungarian vocational schools English in tourism is mainly learnt in an ESP context, only would-be tour guides acquire the discourse of sightseeing in content-based sheltered instruction. To ensure that learners acquire the special tourism terminology, the topic-based approach with the expansion of the relevant core vocabulary proved to be the most effective way of teaching English of tourism in my teaching practice. In addition to course books the use authentic materials including Internet sources, brochures and catalogues help create lifelike communicative situations, and the writing tasks are related to various, mainly problematic

issues of tourism. The use of authentic materials and computer assisted language learning promote that students become familiar with the specific genres of tourism, such as correspondence with customers or tour operators, dealing with hotel guests, or the discourse of sightseeing. Authenticity involves not only providing materials, but also creating situations that are related to learners' target activities and professional objectives. Students get more involved in special language learning when they bring tourism related texts to classes or search for Internet sources, discuss the content, work out the meaning of the new vocabulary items, and share them with their peers.

1.4. Concluding remarks

The development, key notions and main aspects of ESP and content-based instruction have been overviewed in the chapter. Needs analysis, a central element of ESP practices, provides framework to define the specific skills, lexis, grammatical structures and communicative practices that learners of a particular profession need to acquire. Both ESP and content-based instruction aim to equip students and professionals with academic, technical, and interpersonal communication skills, and, to enhance their work-related and academic background knowledge. In ESP, based on targeted needs analysis and syllabus design, language instruction is dominant, while in content-based instruction language is the medium of acquiring the content.

According to the core principle of CBI, a second language can be learnt more successfully when language is used as a means of acquiring information rather than as a learning objective itself (Richards & Rodgers, 2001, p. 209). In CBI language can be considered a scaffolding tool, when it is acquired simultaneously in context meaning is more comprehensive; the CLIL context through incidental learning gives insight into the morphological, syntactic, collocational and pragmatic features of new vocabulary. Encountering new words in context enhances learners' language skills and vocabulary knowledge, and activates their background knowledge by linking new concepts to existing ones. Unlike in other approaches to language teaching, in CBI fluency is more important than accuracy, thus, content-based learning can commence with teaching vocabulary relevant to a topic, instead of teaching grammatical structures (Motuziene, 2013, p. 332).

The following chapter sheds light on some of the major points of vocabulary teaching and learning. It concerns with different aspects of word knowledge, discusses the key concepts of formulaic language and collocations, as well as the main considerations of ESP and content-based vocabulary instruction.

CHAPTER TWO

KEY ASPECTS AND CHALLENGES OF VOCABULARY LEARNING AND TEACHING

2.1. Introduction

Over the past few decades, a gradual but significant change has taken place in second language (L2) teaching and learning; opposed to traditional approaches that gave priority to grammatical competence, in the 1970s communicative language teaching (CLT) introduced the concept of communicative competence, emphasising that language is used for meaningful communicative purposes. Communicative competence, a term coined by Hymes (1972) refers to language users' lexical and grammatical knowledge, as well as to their social knowledge that help them communicate appropriately in different contexts. With the expansion of CLT that focuses on learners' fluency (Richards, 2006, p. 1), lexis has gradually become recognized as crucial in the second language acquisition (SLA) process.

Owing to the lexical approach, a method of teaching foreign languages in the 1990s that exploited frequently occurring lexical phrases (Lewis, 1993, p. 51), greater emphasis has been laid on vocabulary acquisition, which resulted in a great deal of studies on learners' individual differences, bilingual lexicon, vocabulary learning strategies and teaching methods (Booij, 2012; Dörnyei, 2005; Ellis, 1997; Gass & Selinker, 2008; Kroll & De Groot, 2005; Meuter, 2009; Nation, 1990, 2001; Schmitt, 1997, 2000). Learning formulaic language, as an essential element of native-like language production, has received increased interest since the 1990s (Lewis, 1993; Nattinger & DeCarrico, 1992), with focus on the acquisition of collocations and the development of collocational competence (Durrant, 2008; Hill, 1999; Hoey, 2005; Kennedy, 1990; Wray, 2002).

Despite the more conducting role of vocabulary in SLA research and the emphasis on the importance of lexical knowledge, the acquisition of a considerable lexicon has remained a challenge for many language learners that seems to be more perplexing and demanding in academic or professional context. The acquisition of vocabulary is often a stumbling block to

students' language learning, due to lack of exposure, effective vocabulary learning strategies, or lack of self-confidence. Since vocabulary acquisition is an incremental process (Scott & Nagy, 2004, p. 108), the main aim of vocabulary instruction is to equip students with effective vocabulary learning strategies and skills to enable them become independent learners. To facilitate vocabulary learning teachers, material writers, and researchers also need to contribute to the learning process. Thereby, teaching and learning academic and technical vocabulary, and the development of learners' lexicon to achieve professional communicative competence have been in the focus of a body of current research (Bowers et. al., 2011; Dugan, 2010; Luka, 2007; Marzano & Pickering, 2006; Moreno, 2011; Zwiers & Crawford, 2011) with a line that is grounded on and integrates the achievements of ESP and corpus-based studies (Boulton *et.al.*, 2012; Bowker & Pearson, 2002; Coxhead, 2012, 2013; Chung & Nation, 2003; Hou, 2014; Gavioli, 2006).

To underscore my investigations on the potential of corpus-based vocabulary development in tourism English, this chapter introduces some basic terms of vocabulary instruction, concerns the key aspects of vocabulary description and acquisition, and presents developments in current vocabulary research. Different approaches to vocabulary knowledge are concerned including the receptive/productive continuum and the access to the mental lexicon. The concept of formulaic language is discussed in brief and the notions of collocation and colligation are clarified, touching upon the role of corpus-assisted research in their investigations. Following the discussion of vocabulary learning strategies, some characteristics of ESP and content-based vocabulary instruction and acquisition are presented.

2. 2. Terms to discuss in vocabulary acquisition

Vocabulary terms do not always seem to be consistent in vocabulary research, thus, the discussion of the issues of vocabulary acquisition and ESP vocabulary instruction need to commence with providing a clearer understanding of vocabulary terminology. The terms *word*, *vocabulary item* and *lexical item* are often used interchangeably to refer to individual words, whereas their counterparts, *words*, *vocabulary, lexicon* and *lexis* are used when referred to a group of words (Gardner, 2013, p. 10). To handle multiword units, the term *lexeme*, also called *lexical unit/phrase* or *lexical item*, was coined. The latter terms can be defined as 'items that function as single units regardless the number of words they contain' (Schmitt, 2000, p. 2).

The concept of word is assigned in its practical orthographic definition as…'any sequence of letters bounded on either side by a space or punctuation mark' (Carter, 1992, p. 4). In Bloomfield's (1983, p. 99) grammatical definition that stresses the basic stability of a 'word', it is the smallest free form that has a meaning on its own. This definition is challenged by Carter (2012, p. 5) from a semantic perspective arguing that it would be more appropriate to define a word as a minimum meaningful unit of a language. His definition presupposes clear relations between single words and the notion of meaning, and allows differentiating the separate meanings of a word. Yet, there is not necessarily a one-to-one correspondence between a meaning and a single word as meanings are often represented by multi-word lexical units (Schmitt, 2000, p. 1). Words that have restricted identifiable meaning, since they indicate grammatical relationships in a phrase or sentence, are either grammatical or functional words (also known as 'functors', form- or empty words, e.g., *you, the, could, by*). Opposed to function words, lexical words, (which are also known as content- or full words e.g., *tree, nice, build, sadly)*, have greater 'lexicality' as they carry semantic information content, and are syntactically structured by grammatical words (Carter, 2012, p. 8).

The broader term, vocabulary, is defined in the *Collins English Dictionary* as all the words contained in a language, and the total number of words known to an individual person in a particular language that involves recognizing and understanding, although not necessarily using these words by that particular person. The term of Greek origin, *lexicon,* generally occurs in two pertinent meanings; (1) a *dictionary*, the collection of all words and phrases of a particular language, and (2) a mental system known and referred to as *mental lexicon* or *mental dictionary*. This mental system contains all words and phrases known by an actual language user and what that person knows of the form, meaning, and syntactic properties of these entries (Gardner, 2013, p. 10). Williams (1994) suggests that clear distinction be made between two notions of lexicon; the Bloomfieldian lexicon 'as a repository of all of a language's idiosyncrasies' and the grammatical lexicon, 'the linguist's theory of the category of linguistic object that we call a word' (Bloomfield, 1942, as cited in Williams, 1994, p. 7).

The most basic form-based concept of vocabulary, *type*, is generally defined as one or more conterminous letter(s) that form a distinct word with different spelling, and are separated by space in writing, and stress boundaries in speech. *Token* indicates the frequency, the number of occurrences of types or word units within a text, a set of data, or discourse (Gardner, 2013, p. 99). As an instance, the quote by Blake "Tiger, tiger burning bright in the forests of the night" contains ten tokens and eight types as the words 'tiger' and 'the' occur twice. In morphology,

lexicography, and corpus linguistics *lemma* is the base form of a set of lexical items, in Francis and Kucera's (1982, p. 1. as cited in Knowles & Don, 2004, p. 70) definition it is 'a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling'. Corpus linguists treat lemmas for practical purposes as a group of inflectional variants of the same headword (Biber *et al*., 1998; Kennedy, 1998; O'Keefe *et al*., 2007; Sinclair, 1991); however, the individual members develop their own meanings and collocations independently (Knowles & Don, 2004, p. 71).

Terms, or terminological units of specialised lexis refer to an entity, or express a concept related to a particular kind of language or a branch of study. Terms represent concepts in an unequivocal way in a specific semantic field, and they are manifestations of specialised, extended meanings of words in general language (Cabré, 1999, p. 35). Alexeeva and Novodranova (2006, p. 25) maintain the stance of traditional terminology; they consider the term as a unit of language as it exhibits all the parameters of a linguistic sign. They make distinction between two different views on defining the notion of term; according to the linguistic approach the term is defined as a special function of the word, whereas in the cognitive view term is regarded as a sign that transfers not only factual information but particular word knowledge on the situation. In their view *term* is not a category of language but a category of thought and knowledge, the result of conceptualisation (2006, p. 28).

2.3. What does knowing a word imply?

2.3.1. Approaches to word knowledge

Scholars of vocabulary research are all in agreement that words can be known at various levels (Graves, 2006, p. 12). Beck, McKeown, and Kucan (2002, p. 10) argue that the knowledge of a word moves along a continuum on which they differentiate five levels:

> ➢ no knowledge
> ➢ general sense, such as knowing positive-negative connotations
> ➢ narrow, context-bound knowledge
> ➢ having knowledge of a word but not being able to recall it to apply in appropriate situations
> ➢ rich, decontextualized knowledge of a word's meaning, its relationship to other words, and its metaphorical uses.

Investigating the nature of lexical knowledge three primary aspects have been specified by applied linguists: (1) *form* including spelling and pronunciation, (2), *meaning*, and (3) the *relationship between words and their neighbours* including polywords, collocations, colligations, and idioms in formulaic language (Gardner, 2013, p. 10). Nation (2001, p. 27) shed light on the different morpho-syntactic, semantic and pragmatic components of this lexical knowledge scheme including (1) *form* (either spoken or written, and word parts), (2) *meaning* (concepts/referents and associations), (3) *position* (referring to grammatical functions, collocations), and (4) *function* (register, frequency, and constraints of use). Celce-Murcia and Larsen-Freeman (1999, p. 4) recommended accepting a formal and functional approach in language teaching with meaning as a separate dimension, while laying equal emphasis on the interconnected parts of 'form', 'meaning', and 'use'.

Concerning the qualitative aspects of word knowledge, Beck *et al.* (2002) adopted Cronbach's division of qualitative dimensions (1942, cited in Beck et. al., 2002, p. 10): (a) *generalization*, knowing the definition of a word, (b) *application*, knowledge of use, (c) *breadth*, the knowledge of multiple meanings, (d) *precision*, knowing the appropriate use of words in different situations, and (5) *availability*, the productive use of a word. Previous tenets drew on both the quantitative and qualitative aspects of word knowledge; Chapelle (1998, p. 39) argued that vocabulary knowledge comprises four intrinsically connected factors: (a) *vocabulary size*, (b) *knowledge of word characteristics*, (c) *organization of lexicon* and (d) *processes of lexical access*. Three dimensions of lexical competence were proposed by Henriksen (1999, p. 303): (a) *partial vs. precise knowledge*, (b) *depth of lexical knowledge*, and (c) *receptive/productive vocabulary use*.

The complexity of what knowing a word implies was emphasised by Nagy and Scott (2000, p. 270) who specified five aspects: *incrementality*, *polysemy*, *multidimensionality*, *interrelatedness*, and *heterogeneity*, extended with a sixth aspect (Scott & Nagy, 2004, p. 108), *understanding the role of definitions, context and word parts* in vocabulary learning. They claimed that knowing a word moves beyond the definition of word meanings, however, definitions, word parts and context supply important information about the meanings of words. Knowing a word implies knowing its literal meaning, connotations, spelling, derivations, collocations, frequency, pronunciation, its morphological and syntactic constructions, and semantic associates as synonyms, antonyms or homonyms (Nagy & Scott, 2000, pp. 274-278). As many words have different meanings depending on the context, another aspect of word knowledge is the awareness of the multiple meanings, literal/denotative and

implicit/connotative meanings of words that are inextricably related to the knowledge of other words (Scott & Nagy, 2004, pp. 108-109).

First language (L1) greatly influences the acquisition of vocabulary and the clarification of word meanings in a second language (L2). Several factors effect L2 vocabulary learning and word knowledge including *word forms* (spoken or written), *word structure* (derivations and inflections), *syntactic patterns*, and *semantic features* of a word. Laufer (1997, pp. 149-153) drew attention to four semantic properties of words; (1) *abstractness*, (2) *specificity and register restrictions*, (3) *idiomaticity*, and (4) *multiple meanings*. Abstract words tend to be more difficult to learn than concrete words, and lexical items frequent in one field may be unusual in another discourse. Another point is that idiomatic expressions are more difficult to understand and learn than their non-idiomatic counterparts, e.g., *walk down the aisle* and *get married*. What makes the acquisition of words truly complicated is the fact that one form has several meanings and one meaning can be represented by different forms.

From a lexical semantics perspective, Cruse (1986, p. 271-285) distinguished four main types of meanings of words and utterances: *propositional*, *expressive*, *presupposed*, and *evoked meaning*. Propositional meaning refers to the relation between a word and its denotation as it is conceived by a speaker, this type of meaning helps to judge whether an utterance is true or false. Expressive meaning relates to the feelings or attitude of a speaker, presupposed meaning arises from the selectional or collocational restrictions of co-occurrence, whereas evoked meaning emerges from register and dialect variations.

Chomsky (interviewed by Andor, 2015) contrasted the knowledge of the meanings of words with the encyclopaedic levels of knowledge. He asserts that knowing all the meanings of words perfectly with limited lexical knowledge leads to having difficulties in the various uses of language. One of the uses of language is communication, yet 'most of our language use is simply construction of thought (p. 147)'. He holds the stance that in secondary use of language like communication, knowing a word is central as it has always been, but shared knowledge is more involved in communication than encyclopaedic knowledge.

2.3.2. The receptive/productive aspects of lexical knowledge

The awareness of words and their meanings is often described as *receptive/ passive* referring to the words used for comprehension in reading and listening, and *productive/active* also called

expressive vocabulary knowledge that involves words used for communication in speaking and writing (Dugan, 2010, p. 6). This distinction separates the breadth and depths of vocabulary knowledge; breadth referring to the size of lexicon, and depths to the degree of lexical knowledge. Corresponding with Beck et al.'s (2002, p. 10) paradigm of word knowledge, Laufer (1997, p. 126) claims that receptive vocabulary does not transfer immediately to productive vocabulary use, as learning a word moves along a continuum with passive/receptive and active/productive word knowledge at two ends. In contrast to this view, Vermeer (2001) questioned whether breadth and depth are independent constructs of vocabulary knowledge: 'Another assumption is that a deeper knowledge of words is the consequence of knowing more words, or that, conversely, the more words someone knows, the finer the networks and the deeper the word knowledge (p. 222)'. Nevertheless, it is widely acknowledged that in a learner's lexicon the size of receptive vocabulary exceeds that of productive vocabulary, and reception precedes production (Ma, 2009, p. 40).

In Anderson and Freebody's (1981, p. 92) definition, the breadth aspect of lexical knowledge refers to 'the number of words for which the person knows at least some of the significant aspects of meaning'. The size of vocabulary is closely associated with reading comprehension; the breadth of vocabulary knowledge is a good predictor of a learner's reading comprehension ability, and the ability to acquire new information from context (Read, 2007, p. 107). However, according to Anderson and Freebody's (1981, p. 96) knowledge hypothesis not only vocabulary size but also the knowledge of the meanings of words and the concept that they represent, which are parts of the reader's background knowledge, contribute to reading comprehension.

Productive vocabulary is defined by Hiebert and Kamil (2005, p. 3) as a set of well-known, familiar, frequently used words that the individual can use in speaking or writing. The dimension of depth has been investigated from three overlapping perspectives in L2 acquisition research to measure the quality of lexical knowledge; (1) *precision of meaning* (the difference between having a vague idea and specific knowledge of a word meaning), (2) *comprehensive word knowledge* (including orthographic, phonological, morpho-syntactic, semantic, collocational and pragmatic characteristics), and (3) *network knowledge* (the incorporation of words into the lexical network in the mental lexicon) (Read, 2004, pp. 211-212). The first approach is based on the assumption that word meaning can be acquired up to the highest level, the second one attempts to integrate the different components of lexical knowledge, whereas the third approach conceptualises the depth of vocabulary knowledge in terms of associations and interconnectedness in the mental lexicon.

Beyond the breadth and depth dimensions, Laufer and Paribakht (1998) differentiated three types of vocabulary knowledge: *passive*, *controlled active*, and *free active*. Understanding the most frequent meanings of words implies *passive knowledge*, *controlled active knowledge* involves recall with the help of a cue, while *free active knowledge* refers to the spontaneous use of words. Laufer and Paribakht's investigations revealed that passive vocabulary knowledge developed faster and it was significantly larger, whereas active vocabulary growth, especially the extension of free active vocabulary, was slower and smaller.

Gairns and Redman (1986, p. 89) argued that the acquisition of lexis can never be complete as it enlarges and deepens throughout lifetime. As breadth and depth are interconnected constructs of a learner's vocabulary knowledge Read (2004, p. 212) and Vermeer (2001, p. 222) advised the parallel development of vocabulary size and quality of lexical knowledge. In terms of vocabulary size predictive positive correlation was found between lexical competence and L2 proficiency (Meara & Jones, 1988; Meara, 1996), and it was confirmed (Zareva et al., 2005; Wang, 2014) that overall L2 proficiency can be predicted based on lexical competence at productive level. However, examining the relations between Hungarian university students' proficiency exam results and lexical knowledge Lehmann (2009) did not find significant correlation between productive vocabulary use and proficiency level. An explanation of the lack of relationship might be that university students' receptive vocabulary size far exceeds that of their productive vocabulary. Nevertheless, it was corroborated that proficiency test results, including reading and listening comprehension, writing, and grammar in use components, depended on the size of receptive vocabulary.

2.3.3. Access to the mental lexicon

The acquisition of an L2, besides several variables including cognitive and affective factors, is influenced by the organisation of the mental lexicon. The mental lexicon is defined by Jarema and Libben (2007, p. 2) as 'a cognitive system that constitutes the capacity for conscious and unconscious lexical activities. Celce-Murcia and Larsen-Freeman (1999, p. 29) regard this mental system as an inventory of words and productive word derivational processes that comprise not only single words but also compounds and multiword units. Lexicalization is the process when new words enter the lexicon (Booij, 2012), and lexical knowledge involves knowing how to combine elements in the mental system to create new lexical units (Gass & Selinker, 2008).

In Chomsky's (1986, p. 31) view the individual mental lexicon, in his terminology I-language or internal language, is more or less the complete imaging and representation of the language. The words in the mental lexicon have a number of relationships; besides fairly stable, idiomatic word combinations there are individual associations among words based on personal experience. Therefore, the mental lexicon can be conceived as a multidimensional web of words with all kinds of connections between those words based on semantic, phonological, and morphological relationships (Booij, 2012, p. 235). A new word entering the mental lexicon is anchored when it has stable associations and interconnected links with previously acquired words, thus, vocabulary acquisition can be regarded as creating associations in a learner's lexicon (Meara, 1997, p. 119). Takac (2008, p. 11) holds the stance that this mental system is organised and structured, so that people can recognise and retrieve the words they need at an astonishing rate from an enormous number of lexical items. A speaker can find a word mentally in less than 200 milliseconds, which is a proof that the mental lexicon is organised in a way that facilitates access and retrieval (Faber & Uson, 1999, p. 16). The organization of the mental lexicon and the quick process of retrieval pose the question of how memory works.

Memory and its subcomponents, short term, long term and working memory, take a significant part in L2 language acquisition (Schmitt, 2000, p. 131). New vocabulary items are first stored in short term memory, where retention is generally not effective when the number of chunks of information exceeds seven, then, after repetition are stored in long term memory (Thornbury, 2002, p. 24). Working memory facilitates the process of lexicalization and retention; the overall capacity of working memory, the working memory span, has a central role in processing and storing information (Dörnyei, 2005, p. 51). Working memory is more than just a passive transitional storage stage that precedes long term memory; it is an independent workspace with cognitive processing functions such as comprehension and production of language (Dörnyei, 2005). In contrast to implicit vocabulary learning explicit learning, which involves memorising a series of details of vocabulary items, makes demands on the working memory and requires more effort and attention (Ellis & Loewen, 2009, p. 3).

Research on the bilingual mental lexicon (Meuter, 2009; Kroll & De Groot, 2005), and on the relationship between memory and second language acquisition (Ellis, 2001; Dörnyei, 2005; Martin & Ellis, 2012) have been directions of interest since the turn of the millennium. Studies on bilingual mental lexicon have resulted significant findings from psycholinguistic and neurolinguistic perspective; evidence supports that there are common conceptual

representations with overlaps across languages (Meuter, 2009; Kroll & De Groot, 2005; Navracsics, 2007).

As words vary on several dimensions such as their concreteness, their morphological, phonological and orthographic complexity, frequency and cognate status (Kroll & De Groot, 2005, p. 15), all these factors affect how easy is to learn a word and how it is represented in the mental lexicon. Similarity in the phonological structure of the words to be learned and the ones in the learner's native language can have positive effects; learning familiar words is generally more successful than when phonologically different words are presented (Kroll & De Groot, 2005, p. 4). Concrete nouns and cognates are more similar in meaning, this way they are easier to learn, in contrast, the representation of abstract nouns is different and largely determined by context. Cognates, as their morphology is similar, are clustered in the lexicon irrespective of languages, whereas non-cognates are lexically distributed (Meuter, 2009, p. 10).

The vocabulary acquisition of a second language (L2) is greatly influenced by first language (L1) vocabulary. In early bilingual vocabulary acquisition, the same way as in L1, the acquisition of word form and meaning is a parallel process, whereas in late foreign language learning the meaning of the new words is already set. In the bilingual lexicon at the stage of fluency, thousands of hours contextualized L2 vocabulary use creates direct connections between L2 word forms and their meanings, while early L2 vocabulary acquisition relies on L1 phonological and conceptual representations and word-concept mappings. L2 vocabulary becomes independent from L1 only after considerable L2 experience; however, rapid and automatic access to the words stored in the bilingual mental lexicon is a prerequisite of fluent communication and reading (Kroll & De Groot, 2005, pp. 9-10).

Navracsics (2007) examined the structure of bilingual lexicon of learners whose L1 or L2 was Hungarian, and investigated whether there are differences between L1 and L2 storages and lexemes. She found that fluent bilinguals use conceptual representations that are shared across their languages, and concluded that the memory of bilinguals contains structures of various types that occur in different proportions, depending on the level of proficiency, word characteristics, learning strategies, context, and age.

2. 4. Formulaic sequences and collocations

2.4.1. Collocations and multiword lexical units

Due to the interest in the syntagmatic aspects of lexicon in the 1980s Nattinger (1980, cited in Carter & McCarthy, 2014, p. 49) proposed that vocabulary teaching should shift from single words to multi-word lexical units that did not seem to fit either traditional notion of grammar nor of vocabulary. Formulaic language generally refers to multi-word phrases that are stored and retrieved holistically, including idioms, set expressions, and collocations. Howarth (1998, p. 42, cited in Cruse, 1986, p. 37) classified multi-word lexical units as (a) *free combinations of words*, (b) *rule-governed word combinations*, and (c) *idioms* 'whose meanings cannot be inferred from its parts'. Howarth (1986) maintained that free word combinations and stable multi-word units such as idioms are predominantly unproblematic for learners, whereas collocations, recurring multi-word combinations, are difficult due to their pervasiveness, arbitrariness and restrictedness. Collocations, a subset of formulaic sequences, form two- or three-word syntagmatic units that can consist of both lexical and grammatical words (Henriksen, 2013, pp. 29-30). A node word or basis has numerous collocates that are neither completely free lexical choices nor fixed either. Consequently, collocations can be defined as phrases that are more restricted than free combinations but less restricted than idioms (Lesniewska, 2006, p. 95). Nevertheless, there are blurred transitional areas between these categories; free combinations can be at the borderline of collocations and collocations can be idiomatic (Cruse, 1986, p. 41).

The notion of collocation, a term introduced by Firth (1957), has been defined in various ways depending on the specific aim of the researcher. The Firthian definition of collocation drew attention to the context-dependent nature of the meaning of co-occurring word patterns: "You shall know a word by the company it keeps" (Firth, 1957, cited in Römer, 2005, p. 12). His notion of meaning differed from the semantic, paradigmatic views (1957, p. 196); 'Meaning by collocation is an abstraction at the syntagmatic level, and is not directly concerned with the conceptual or ideal approach to the meaning of words'. Sinclair detected, drawing up the basic idea of corpus linguistics, that a single word carries only meaning through several words in a sequence (1991, p. 170), he regarded collocations as co-occurrence of word forms that tend to have different sets of collocates. In Benson's (1990, p. 2.) definition collocations are 'arbitrary and recurrent word combinations'. Opposed to this view, the mutual expectancy of juxtaposed words is emphasized in the influential formulation of collocation proposed by Hoey: 'the

relationship a lexical item has with items that appear with greater than random probability in its textual context' (1991, p. 7). Hill's (1999, p. 3.) description of collocation corresponds with that of Hoey (1991); 'the way words occur together in predictable combinations'.

Collocations have also been studied from a syntactic perspective in relation to grammatical patterning or colligation, in the typical syntactic context in which they occur. According to the lexis-grammar approach (Hoey, 2000; Sinclair, 1991) collocations occur at the intersection of lexis and syntax, thus, clear distinction cannot be drawn between them. A re-stated definition of collocation given by Gabrielatos (2018, p. 244) claims that collocations are not purely lexical, but they are influenced by grammar, and their semantic preferences are lexicogrammatical features, thus, every instance of language use is lexicogrammatical.

The definitions approach collocations from three different perspectives: (1) *semantic/syntactic* type of word combinations that is traditional in lexicology, (2) *statistical/textual* that identifies collocations in terms of frequency of co-occurrence, and (3) *discoursal/rhetorical* that defines collocations in terms of performance (Gledhill, 2000, pp. 7-13). Narrowing down the scope of the definitions of collocations in Antle's division (2013, p. 346) they tend to encircle two main groups: *statistical* and *phraseological*.

The statistical approach to defining collocations is pursued by corpus linguists (Durrant, 2008, 2009, 2010; Hoey, 2000, 2005; Sinclair, 1991; Stubbs, 2002) grounded on corpus-driven collocation research. In Sinclair's (1991, p. 170) statistical/textual definition: 'collocation is the occurrence of two or more words within a short space of each other in a text'. Sinclair *et al.* (2004, xiii) provide a statistical working definition of collocations: 'the more-frequent-than-average co-occurrence of two lexical items within five words of a text'. Collocations identified and extracted with corpus-based methods have been termed in various ways such as *lexical bundles, clusters*, *recurrent word combinations* or *collocational networks*.

Sinclair's notion of collocations was extended by Durrant (2009, p. 158) who defined collocations as 'sets of two or more words which appear together more frequently than their individual frequencies would lead us to expect'. From a phraseological perspective, according to Wouden's (2002, p. 5) semantic/syntactic definition, collocation is 'idiosyncratic restriction on the combinability of lexical items'. Gledhill (2000, p. 13) approached the scope of collocation from discoursal/rhetorical perspective; '*the get the sack, to be fired* can be contrasted stylistically with less marked expressions as *to be dismissed*, or *to lose one's job*'.

The relations of words can be (1) *syntactic* such as verb-object or (2) *lexical semantic* relation as synonymy or antonymy, or (3) *pragmatic* relation as in a conventional unit of expression 'You're welcome'. For this reason, not only lexical and syntactic, but also semantic-pragmatic constraints play a role in a speaker's lexical co-selection (Bartsch, 2004, p. 26). Opposed to the Firthian (1957, p. 19) view where grammar and lexis operate at separate levels, Sinclair (1991, pp. 109-115) put forward his complimentary principles to explain how lexical selection is governed and how meaning arises from texts: the *idiom-* and the *open-choice principles*. According to the idiom principle that is central in language production and comprehension, language users select from a set of semi-preconstructed phrases that constitute 'single lexical choices even though they might appear to be analysable into segments' (Sinclair, 1991, p. 110). That is, the idiom principle introduces a lexical restriction besides grammatical ones. The idiom principle is contrasted with the open-choice, also known as slot-and-fill principle, a way of seeing the language text as a complex choice of lexical items to fill a vacant place, and the only constraints are syntactic or pragmatic.

Sinclair (1991, p. 115) considers collocations as illustrations and manifestations of the idiom principle. Barnbrook (2007) reassessed the description of collocation given by Sinclair contrasting the open-choice and idiom principles. As the open-choice principle does not provide enough restraints on consecutive lexical choices, the idiom-principle is suggested to account for meaning; the nature and co-occurrence of collocations are 'at least as important as grammar in the explanation of how meaning arises in text' (Sinclair, 1991. p. 112).

2.4.2. The nature of collocations and colligations

Ellis (1996, pp. 91-93) posited that 'language acquisition is essentially sequence learning that starts with phonological sequencing', and in the same way learners' long-term knowledge of lexical phrases provides basis for the acquisition of grammar, hence his term 'chunking' refers to the general process of SLA. Considering that learners differ in their sequencing ability, comprehension and fluent production involves the acquisition of memorised sequences of language. Accordingly, vocabulary is often learned in lexical phrases comprising several words (Sinclair, 1991, p. 170), and these lexical bundles are retrieved automatically as whole units from the mental lexicon (Wray, 2000, p. 465). Eventually, these multi-word units are analysed into individual words plus grammar through a segmentation process (Ellis, 1997, p. 229).

The frequent co-occurrence of two words as linguistic input helps to create associations in long-term memory; when the same lexical units are encountered in a subsequent input they are perceived by learners as units, and then further reinforced as long-term representations (Durrant, 2008, p. 59). Cognitive psychology suggests that our brain works with a limited attentional capacity when we communicate, and with the help of lexical chunks communication requires less effort (Mayo, 2007, p. 119). The idea that formulaic language underlies fluent language production is supported by research on automaticity in language processing (Durrant, 2008; Ellis, 2008, 2012; Wray, 2000, 2002), emphasising the role of formulaic chunks in fast and effective processing and retrieval.

Prototypical collocations meet three criteria: (1) *non-compositionality*, the meaning of the collocation is not made up of the meanings of the parts, e. g. *white lies* (2) *non-substitutability*, and (3) *non-modifiability* (Dickinson, 2013). Thornbury (2002, p. 221) differentiated two basic categories of collocations based on the parts of speech their constituents belong to: *lexical* and *grammatical* collocations. Lexical collocations consist of diverse combinations of nouns, verbs, adjectives and adverbs; while grammatical collocations contain a noun, verb, or adjective combined with a preposition or a grammatical structure such as a clause or infinitive.

In Baker's view (1992, p. 63) collocations are fairly flexible patterns of language with several variations in form. She classified collocations based on the range of vocabulary items that constitute a collocation as *open* and *restricted collocations*. In open collocations the node that has a general meaning occurs with numerous collocates e.g., *run a business, company, marathon* etc., whereas a node with specific meaning has a more restricted range of collocates e. g. *rancid butter/oil* (1992, pp. 49-52).

Hill's (2000, pp. 63-64) taxonomy of collocations is based on the strength of cohesion and the restrictedness of word combinations:

➢ unique collocations: e.g., *shrug your shoulders* (the verb is not used with other nouns)
➢ strong collocations: e.g., *torrential rain* (torrential is generally used with rain)
➢ medium-strength collocations: e.g., *throw a party* (neither free nor completely fixed)
➢ weak collocations: e.g., *short hair* (these word combinations are free)

Collocations occur in different forms (Dickinson, 2013, p. 9), for instance: (1) light verbs e.g., *do, have, make, take* that convey little meaning but must be the appropriate ones: e.g., *make a*

reservation, (2) phrasal verbs composed of a main word and a particle: e.g., *to take off,* or (3) terminological expressions, terms that form a unit: e.g., *inbound/outbound tourism*.

The collocation of a node word with a particular grammatical class of words is generally referred to as *colligation* (McEnery et. al., 2006, p. 82). Whereas collocations entail the lexical associations of words, colligations refer to their syntactic relations in context; a lexical item tends to co-occur with certain grammatical categories in a particular textual position (Cheng, 2014, p. 2). In Hoey's (2000, p. 234) interpretation 'colligation can be defined as the grammatical company a word keeps and the position it prefers'. He (2005, p. 43) describes the textual position and function of words based on the assumption that words prime each other: 'a lexical item may be primed to co-occur with another lexical item, so also it may be primed to occur in or with a particular grammatical function'. Stubbs defined colligations as 'the relation between content and function words, and between words and grammatical categories' (2002, p. 238). Summarizing, colligations refer to the formation of grammatical structures such as *that*-clauses or infinitive complements, whereas collocations denote constrained lexical choices (Bartsch, 2004, p. 31).

When form shifts to meaning there are always semantic relations between the node and the collocates and among the collocates themselves. These semantic relations, *semantic prosody* and *semantic preference*, are two independent collocational meanings. The interaction referred to as '*semantic prosody'* can be identified as associating evaluative positive and negative or neutral feelings to meaning, that is, the semantic properties of a lexical item determine its collocates (Stubbs, 2002, p. 225). According to Louw (2000, p. 58), 'the primary function of sematic prosody is to express speaker/writer attitude or evaluation'. Stubbs (2001, p. 65) defined the notion of *semantic preference* as 'the relation not between individual words, but between a lemma and a word form, and a set of lexically related words', and proposed that 'the meaning arising from the common semantic features of the collocations of a given node word can be referred to as semantic preference' (2002, p. 225). Nevertheless, as a response to Sinclair's (1996) claim that from a pragmatic perspective semantic prosody is the discourse function of a sequence, Stubbs (2001) re-assessed the concept of semantic prosody and re-named it as 'discourse prosody'.

As claimed by the theory of lexical priming, the semantic properties, the associated meanings determine the collocates, thereby each word is primed to be used with others (Hoey, 2005, pp. 5-8; Stubbs, 2002, p. 225). Lexical priming involves the subconscious noticing of word

meanings, grammatical patterns, style, and register so that the speaker, without realizing, reproduce them in their own speech and writing (Hoey, 2009, p.1). Nevertheless, Pace-Sigge (2013, p. 153) maintains that lexical priming is not a linguistic but a psychological concept.

Tutin (2008, pp. 1-8) extended the definition of collocation by examining the binary status of collocations. Based on the definition that collocations are associations of two lexical units, she argues that collocations are basically argument-predicate structures, where the predicate is the collocate and the argument is the base as in *heavy* (predicate) *rain* (argument). She supports the view that most collocational clusters can be considered as combinations of binary lexical collocations, and proposes the term 'recursive collocations' when collocations are inserted into other collocations as *freshly baked bread*, for instance, where the collocate itself is a collocation.

Durrant and Doherty (2010, pp. 125-155) investigated whether binary high-frequency collocations are psychologically real, drawing on the thesis of collocational priming. They support Hoey's (2005) ideas that high-frequency collocates mentally prime or predict each other, and frequently co-occurring words might have some independent psychological associations and representations in speakers' mind. They tested to what extent collocational priming exists by focussing on simple 'automatic' associative processes, which they thought to best reflect the organization of the mental lexicon. In addition, they examined as to whether high-frequency collocations identified in a corpus are similar to those represented in the minds of proficient speakers. Associative collocational priming was found to exist, confirming that frequency of occurrence does indicate psychological reality, and, suggesting that frequency-based methods are valid for the identification of collocations that should be targets for language learning.

2.4.3. Pedagogical implications of collocations

Collocations have been considered important aspects of a language learner's word knowledge since the 1990s (Howarth, 1998; Nesselhauf, 2003), as collocational knowledge not only improves accuracy, but it also promotes fluency and the development of pragmatic skills. According to Hill (2000), 'within the mental lexicon, collocation is the most powerful force in the creation and comprehension of all naturally occurring texts' (p. 49). Nevertheless, while idiomaticity facilitates fluent production for native speakers effortlessly, it is an obstacle to

surmount for language learners as their collocational knowledge is limited (Hill, 1999, p. 4). Stubbs (2001, p. 73) highlighted the correlation between collocational proficiency and nativeness; 'Native speakers' unconscious knowledge of collocations is an essential component of their idiomatic and fluent language use, and an important part of their communicative competence'. Apparently, there is a clear-cut bond between collocations and naturalness; however, the word combinations of an L2 can be acquired by non-native speakers.

As Wray (2002/b) accentuated, learning formulaic language is possible through conscious effort, yet, the greatest fallacy of learning collocations is manifested in the erroneous use of words. Korosadowicz-Struzynska (1980) observed that 'errors in the use of word collocations surely add to the foreign flavour in the learners' speech and writing, and along with his faulty pronunciation they are the strongest markers of an accent' (p. 115). Nesselhauf (2005) found that the most frequent types of error of German EFL learners involves the erroneous use of verbs in verb plus noun combinations that are likely to emerge from L1 transfer. De Cock (2003) also detected that French learners misapplied English sequences that have a French congruent form used in a different way, and used L1-based idiosyncratic combinations like *according to me*.

Among the lexical problems of language learners, collocational errors are the most frequent as the most salient markers of non-nativeness. Learners tend to depend on the collocational patterns of their L1; word-by-word errors often occur in learners' L2 performance due to the shortage of ready-made chunks in their vocabulary (Willis, 2003, p. 160). The reason why collocations are generally problematic for language learners is the fact that their collocational competence, the knowledge of lexical phrases that enables them to use language fluently (Daskalovska, 2009, p. 1), does not improve at the same rate as their lexical knowledge (Lesniewska, 2006, p. 95). Collocations are semantically transparent, thus, when learners encounter collocations for the first time in a foreign language, it is likely that they find them comprehensible. Idioms, on the other hand, are often figurative and opaque, this way attract learners' attention as they are perceived as salient that promotes retention (Lesniewska, 2006, p. 98).

The most problematic collocations are at the borderline of free and restricted phrases; a common type of error is the overextension of words that results in awkward collocations. An important characteristic of advanced learners' L2 performance is that deviations from native-speaker norms are often implicit errors as they do not violate the rules of L2 morphology, syntax or

semantics; only the cumulative use of certain phrases gives the impression of non-nativeness (Lesniewska, 2006, 96-97). As an 'avoidance strategy' learners often rely on 'safe bets', generally collocations that have direct translation equivalents in L1, or use 'all-purpose' phrases that can be easily extracted from the mental lexicon (Lesniewska, 2006, pp. 96-98). A frequent violation of restrictions on word selection is the production of erroneous collocations using synonyms that are not interchangeable (Phoocharoensil, 2014, p. 2535).

Durrant and Schmitt (2010, p. 11) argue that L2 learners' difficulties with acquiring collocations are due to insufficient exposure to create associative links between constituents. Gyllstad (2007) also maintained that collocational knowledge can be acquired through high frequency exposure either through formal language learning activities in classroom settings, or through incidental learning outside the classroom, such as reading printed materials, watching television programs, listening to radio, or communicating through social networking sites.

Kennedy (1990) holds the stance that it is collocations where grammar and vocabulary meet. The lexico-grammatical nature of collocations can be an explanation of why learners have problems with producing idiomatically correct language. As discussed above, the lack of collocational competence leads to lexical and grammatical failure as learners tend to use word-by-word complicated phrases instead of collocations, relying on their L1 to compensate for their deficiency in an L2. Consequently, the acquisition of appropriate collocational patterns is crucial for language learners to ease language production; to speak and write in natural and accurate way (Barnbrook, et al., 2013, p. 120). Collocation grids and semantic matrices, semantic mappings, are useful ways of presenting and practicing relations between node words and their collocates, and of helping to avoid the use of awkward word combinations (Gairns & Redman, 1986, pp. 37-38).

Various other means of vocabulary learning activities, including corpus-driven tasks such as gap-filling, multiple-matching based on concordancing, guessing headwords that collocate with given words, or error spotting, facilitate the expansion of learners' collocational knowledge, this way fostering their overall language development. The high-frequency lexical chunks of corpora promote retention; the more lexical patterns observed in authentic context the more likely to retain them in short-term and long-term memory. The importance of collocations has also been recognised in lexicography; as the selection of dictionary entries is based on corpus-driven frequency analysis, dictionaries inform users on the collocational behaviour of words enhancing their awareness of word combinations. Books entitled *The words you need* and *More*

words you need aim to activate learners' collocational sense with the help of collocational grids and semantic matrices, whereas *English collocations in use* and *Using collocations for natural English* present collocations in typical contexts that focus on a topic, providing numerous activities to practice them.

To raise learners' collocational awareness is a challenging task; both teachers and learners agree that formulaic expressions can be more difficult to understand and learn than their non-idiomatic counterparts, as it frequently happens in the case of phrasal verbs. Durrant (2008) observed that enhanced input in the form of repeated exposure to collocation lists can highly increase vocabulary and facilitate retention, thus, learners need to be encouraged to exploit the benefits of extensive reading and explore collocations by examining concordance lines. Hill (2000) drew teachers' attention that they should raise students' awareness of collocations by noticing, as well as encourage learner autonomy, as the development of collocational competence is a long-term process. The improvement of all four skills promotes the acquisition of collocations; reading and listening accommodate noticing and observation, while speaking and writing activities help practice and facilitate retention. Teachers should start to improve learners' collocational knowledge with presenting authentic texts, dividing language into meaningful chunks, and make use of fluency practice to keep strings of words in students' short-term memory.

2.5. Vocabulary learning strategies

The enhancement of learners' word consciousness and motivation, their awareness of words and interest in them so that they recognize and reflect on various features of language, are significant components of vocabulary instruction (Graves, 2006, p. 37). Graves (2006, pp. 4-8) offers a four-component scheme for effective vocabulary development; teachers need to (1) provide *varied and meaningful language learning experiences*, (2) *teach individual words besides incidental learning* as students benefit from systematic direct instruction of words, (3) *teach word-learning strategies*, and (4) *foster word consciousness* to maintain students' interest in learning new words.

Vocabulary learning strategies are effective tools for increasing vocabulary size and deepening lexical knowledge. Oxford (1990, p. 4) considers the strategies that learners apply as 'aids to the acquisition, storage, retrieval and use of information'. Ellis (1995, cited in Takac, 2008, p.

49) advises that distinction be made between *language learning strategies* and *language use strategies*; learning strategies are applied when maintaining contact with learning material (classroom tasks, assignments) and storing that material in memory, whereas language use strategies include retrieval, rehearsal, and cover strategies, the latter are special types of compensation strategies. Vocabulary learning strategies (VLS), sub-categories of learning strategies, help learners understand the meaning of words and lexical chunks, and facilitate learning, retaining, and recalling them.

Oxford's (1990, p. 8) taxonomy of VLS includes two major categories; *direct strategies* (memory, cognitive, compensation), and *indirect strategies* (metacognitive, affective and social). Drawing on Oxford's categorization in Schmitt' scheme (1997, pp. 199-227) two broad types of vocabulary learning strategies are distinguished, *discovery* and *consolidation strategies*. The former includes determination and social strategies, the latter social, memory, cognitive and metacognitive strategies. Discovery strategies involve the different processes to infer word meanings and work out their usage such as guessing from context, and asking peers or teachers for meaning. Consolidation strategies entail processes necessary for rehearsing and stabilizing the information stored in memory, so that learners could recall them in subsequent language use *e.g.*, keeping vocabulary notebook, interaction with native speakers, or using semantic maps.

Learning new words in context is often an effective way to increase L2 vocabulary size (Hulstijn, 1992; Nagy, Herman & Anderson, 1985; Pitts, White & Krashen, 1989) as first language learners acquire most of their vocabulary in indirect or implicit way (Nation et al, 1997, p. 11). However, Hulstijn (2001, p. 275) argues that incidental and intentional learning should be seen as complimentary activities. In accordance with Hulstijn's view, Schmitt (2008, p. 329) maintains that vocabulary instruction needs to include both an explicit, intentional learning component, and a component that optimizes exposure and incidental learning. To integrate intentional and incidental vocabulary teaching Nation (2001) suggested drawing upon the four learning strands; meaning-focused input, meaning-focused output, language-focused learning, and fluency development.

The complex, often multiple meanings of words need to be understood in the context of other words in a text, or discourse (Hiebert & Kamil, 2005, p. 1). The acquisition of word meaning/s is a gradual process involving successive encounters in various contexts; it can proceed explicitly (consciously) with teacher's aid and explanation, or implicitly (subconsciously)

through exposure to language input as extensive reading (Donesch-Jezo, 2013, p. 13). To infer word meanings Graves (2006, p. 23) advises three vocabulary learning strategies: (1) use of context to puzzle out the meaning of unknown words, (2) teaching students to use word parts to elaborate word meaning, and (3) teaching students to use dictionaries and thesauri. Encounters with new words in different contexts reveal several aspects of language and vocabulary such as word structure, form, collocations, word class, semantic relations with other words, and to a certain extent, meaning. Word knowledge, strategic knowledge, and linguistic knowledge contribute to the disambiguation of word meaning through context-based inferences (Ellis, 1997, p. 230).

Learners can boost their lexicon 'if they are exposed to sufficient amount of comprehensible input' (Takac, 2008, p. 17), thus reading is an ideal source for learning new vocabulary items. However, according to Takac (2008, p. 17) 'mere exposure during reading does not guarantee a rapid vocabulary growth'; learners must possess strategic knowledge that enables them 'to turn the incidental learning to an explicit learning process'. Vocabulary development and the development of reading comprehension skills are complimentary processes; vocabulary supports and increases comprehension, and learners acquire new words through understanding what they are reading.

Two approaches to reading texts in ESP are of prime importance; in one of them the text is regarded as a vehicle of information from which extracting the gist quickly is more important than immersing in details, applying this method, learners first process the language and then link them to previous knowledge. The second approach to teaching reading on ESP courses presupposes that purposive reading requires language and language skills; less successful learners had a fragmented approach to texts, while successful learners inferred their overall meaning (Bojovic, 2010, p. 2).

Nation (1990, pp. 247-249) argues that the applied learning strategies are in close connection with the frequency of words, and vocabulary should be viewed from a cost/benefit viewpoint. The 'cost' of teaching high frequency words directly is justified by the results, whereas low frequency words generally do not justify the merits of explicit teaching time, thus indirect/implicit strategies such as guessing from context through extensive listening and reading, using mnemonic techniques and word parts (inflections, affixes), or observing concordances can help students to acquire them.

2.6. ESP and content-based vocabulary teaching and learning

2.6.1. Special vocabulary in professional communication

The enhancement of learners' communicative skills in occupational settings, including the extension of special lexis, is a major concern in special language teaching and learning. The use of technical jargon, special vocabulary with supposed professional background knowledge, is an expressive marker of professional communicative competence (Havril, 2009, pp. 74-75). As CBI curricula are organized on a subject matter, in CBI courses grammar and vocabulary are presented through academic or professional contextualized instruction (Stoller, 2002, p. 125). For learning a subject in a second language (L2) learners need not only technical vocabulary but also basic L2 skills, academic L2 and metacognitive skills to be able to speak, write, read on a wide range of topics, follow and engage in interactions using appropriate grammar and discourse (Ardeo, 2013. p. 27).

Research conducted on the knowledge of domain-specific technical vocabulary of professions and disciplines has been increasingly in the focus of teaching and learning ESP in the past few decades (Chang & Nation, 2004; Coxhead, 2012, 2013; Flowerdew, 2012; Georgieva, 2010; Nation, 2001). Yet, although technical vocabulary is crucial for ESP learners, studies on approaches for defining technical terms and on proving their effectiveness analyse segments and different aspects of technical vocabulary, they do not provide comprehensive categorization or definition (Csomay & Petrovic, 2012; Chujo, Utijama & Oghigian, 2006; Chung & Nation, 2003, 2004; Fraser, 2005; Kwary, 2011; Nation, 2001; Peters & Fernandez, 2013; Sznajder, 2010).

Defining specialized lexis is an intricate task, as we need to consider whether words that are closely related to a subject are specialized, or only those words that are unique in that field. From among the several approaches to the identification of technical vocabulary one frequently used approach is the selection of lexical items based on the intuition of an expert (Chung & Nation, 2003). The selection can be conducted in three ways; (1) using a rating scale (Baker, 1988), (2) using technical dictionaries compiled by subject specialists (Nation, 2001), or (3) by making use of the words marked and considered to be important by the most relevant specialist, the writer of the text (Flowerdew, 1992). Another approach is the corpus-based comparison of frequently occurring words in a technical text with those in different corpora; evidently,

technical words are much more frequent in specialised corpora (Sznajder, 2010; Sutarsyah *et al*, 1994).

Chang and Nation (2004) developed a semantically-based four-point rating scale for the corpus-assisted identification of *high-frequency words*, *academic*, *technical* and *low-frequency words* in specialized texts of academic course books. Their investigations showed that one-third or at least one-fifth of the words in the course books were technical that were closely or exclusively related to their field of study. These figures exceeded the 5% that Nation (2001, p. 12) previously estimated, and highlight the challenge that learners need to acquire vocabulary beyond the high-frequency and sub-technical groups of words. Adopting this rating scale methodology, Fraser (2005) found that many of the most frequent technical terms are crypto technical, that is, common words with a 'hidden' technical meaning.

Special vocabulary is referred to in various ways: 'terminological words, also 'terms' and 'terminology' (Becka, 1972), 'technical vocabulary' (Nation, 2001), 'specialized lexis' (Baker, 1988) or 'specialized vocabulary' (Kennedy & Bolitho, 1984, cited in Chang & Nation, 2004, p. 252). According to Nation (2014, p. 198) special vocabulary includes lexical items that are "recognizably specific to a particular discipline, field or topic". In Marzá's (2009, p. 48) stance "terminology is the most evident, visible and easily 'palpable' characteristic of specialised discourses… it is a fundamental element of specialised languages".

Domain-specific technical vocabulary of professions and disciplines is divided into three basic categories based on their semantic ambiguity by pioneer ESP theorists (Robinson, 1991; Hutchinson & Waters, 1987; Dudley-Evans, 1998); *technical*, *sub-technical* and *general vocabulary* of frequent use in a related field. Technical vocabulary in Rea's definition (2009, p. 162) consists of 'content words whose meaning is restricted to the specific subject, characterises the specific language as an individual area of the global language and constitutes the terminology of the domain'. Technical words (1) are monosemic and are not used out of their specific domain whereas semi-or sub-technical vocabulary (2) includes lexical units which are polysemic general words that acquired new meanings in a specific context. The most numerous group, general vocabulary of frequent use in a specific area (3) are at the boundaries of general and special lexis without losing their original meaning (Alcaraz, 2000, pp. 42-44, quoted in Carbajosa, 2012). Widdowson argues (1983, pp. 92-94) that general or 'procedural' vocabulary as a strategic resource supports the domain-specific vocabulary of texts and helps learners acquire technical vocabulary.

Baker (1988, pp. 91-92) pointed out that the real difficulty with understanding scientific/technical texts is due to the area of vocabulary generally referred to as sub-technical vocabulary, as lexical problems hinder comprehension. Sub-technical or semi-technical vocabulary is defined by Rea (2009, p. 162) as 'general content words whose meaning becomes specialised in a domain but it is understandable from its meaning in a general context'. Sub-technical vocabulary comprises lexical items that are neither specific to a certain field or knowledge, nor general everyday words which are used distinctively in specialized texts, as 'occur', for instance. Cummins (2002, pp. 2-5) shares Baker's (1988) view that sub-technical vocabulary is a barrier to the comprehension of specialized texts, in addition, subject-specific texts generally include longer and more complex sentences than spoken English. Investigating the relationship between reading comprehension and text coverage of specialized vocabulary, Alcantar and Scholfield (2010, pp.1-13) found empirical evidence that for the comprehension of tourism texts it is more essential to know the most distinctively frequent words of tourism terminology than general high frequency words.

Lexicographers, course book writers as well as ESP teachers need to be aware of what lexical items can be considered technical vocabulary when creating specific dictionaries, syllabi and course materials. Recent studies on designating the specific lexis investigate different ESP domains; nevertheless, the applied methods, techniques and procedures can be implemented in determining and encompassing tourism vocabulary as well. Mixed-method research (Peters & Fernandez, 2013) on the lexical needs of ESP students unfolded that learners' lexical searches were targeted more often to the shared terminology of sciences and academic discourse than to merely profession-related words. The technical uses of everyday words were the most challenging for learners; it is difficult to identify meaning when the terms of the target language and the L1 are not congruent.

Corpus-based research on determining and encompassing ESP vocabulary conducted by Kwary (2011) draws on Nation's (2001, pp. 11-13) distinction of four kinds of vocabulary; high-frequency, academic, technical and low-frequency words. Applying Nation's (2011, pp. 11-13) categorization Kwary examined four corpus-based methods that can be effective in determining technical vocabulary; *vocabulary classification*, *keyword analysis*, *term extraction* and *systematic classification*, that is, comparing lexis with ESP dictionaries. Realizing the problems with these methods such as not identified word classes, vague results of cluster analysis, the limited number of words in a cluster and the time-consuming process of systematic

classification, a so-called hybrid method that combines the keyword analysis method and systematic classification was proposed to eliminate these drawbacks (Kwary, 2011).

Chujo et.al (2006) carried out research to determine the most meaningful technical vocabulary extracted from a tourism corpus applying statistical measures, and found that different statistical measures tend to extract lexical items corresponding to a certain grade level or textbook coverage. In a corpus-assisted study Sznajder (2010) analysed the metaphoric-based technical and semi-technical terms of a business English course book by contrasting them with the metaphors identified in a corpus of business periodicals and journal articles. The findings suggested that course books might contain infrequent figurative multi-word items, thus the selection of metaphorical technical terms and collocational patterns should be based on corpus evidence.

2.6.2. ESP vocabulary teaching and learning

As teaching and learning specialised vocabulary/terminology is an integral and important part of ESP lessons, the topics and syllabi that encompass the special lexis need to be congruent with course objectives (Georgieva, 2010, pp. 1; 11). Besides general vocabulary, learners have to learn low-frequency, germane, subject-specific words and phrases as well as academic vocabulary to meet the communicative needs of their future job (Ardeo, 2013. p. 28). Learners acquire new vocabulary items more effectively when the contextual clues help convey word meaning, and the new lexis is practised, thus retained, in meaningful professional context (Larson-Freeman, 2000, p. 142). "ESP course plans, vocabulary teaching programs need to expand learners' potential in extracting meaning and to develop the knowledge of the jargon of the genre" (Riahipour & Saba, 2012, p. 1258). Liuolienė and Metiunienė (2013, p. 51) recommended the use of innovative and creative methods in teaching new technical vocabulary; authentic, real-life texts that are not written for pedagogic purposes, interactive vocabulary activities on the Internet, students' presentations, project-based teaching and role-plays of professional situations. Investigating the potentials of the integration of specialized corpora in a CLIL course Hou (2014, pp. 34-35) advises the use of core vocabulary lists as supplementary materials. Adapting Chung and Nation's (2003) ideas Hou suggests that learners should be helped with recognizing specialized words, interpreting definitions, learning phrases and word parts, relating senses to a core meaning, and with acquiring vocabulary learning strategies.

Learning strategies are effective ways of acquiring competence when incorporated to ESP as well (Duenas, 2004, p. 78) as they enable learners to take more control and responsibility of

their learning, this way enhance learner autonomy (Wanpen *et. al*, 2013, p. 313). As Williams (1985, p. 121) remarks a "typical two- or three-hour a week ESP course would not have the time or space necessary to teach all of the specialized vocabulary students would need to learn". To find a solution he recommended five vocabulary learning strategies to be applied in ESP: (1) *inferring from context*, (2) *identifying lexical families*, (3) *unfolding nominal compounds*, (4) *searching synonyms*, and (5) *word analysis* (1985, p. 122).

Examining university students' strategy use and academic vocabulary knowledge, meaningful relation was found between using metacognitive, cognitive, affective, compensatory and social strategies, and vocabulary knowledge (Bijani *et. al*., 2014, pp. 1-16). The most effective strategies that helped learners increase their vocabulary were cognitive and metacognitive strategies, that is, learners were aware of their strengths and learning style, and consciously monitored and changed their learning behaviours. Research on Hungarian university and high school students' vocabulary learning strategies (Dóczi, 2011, p. 138-153) revealed that social and metacognitive strategies were less often used, the number of strategies increased with the time spent studying the language, and the most popular strategy applied was guessing from context, followed by the use of monolingual dictionaries.

A mixed-method study on engineering students' technical vocabulary learning strategies (Wanpen, Sonkoontod & Nonkukhetkhong, 2013, p. 312-320) attested that students who had vocational educational background were more proficient and had larger technical vocabulary than those who received general education. However, the students in vocational stream applied vocabulary learning strategies less frequently than students in the general education stream, yet, in both groups metacognitive strategies were the most frequently used strategies.

In the contextualized language learning of ESP and CBA courses, besides intentional learning, new vocabulary items can be acquired applying implicit strategies including extensive reading, exposure to natural listening input *e.g*., news items, films and TV programmes, and meaning-focused communicative interactions (Mukoroli, 2011, p. 24). As ESP vocabulary teaching generally focuses on explicit vocabulary teaching and learning, Csomay and Petrovic (2012) detected legal technical terms in a corpus of discipline related movies and TV shows, and found that with the recurring lexical patterns they facilitate incidental ESP vocabulary learning, thus could be used effectively to enhance technical lexicon.

2.7.   Concluding remarks

This chapter presented some major aspects of vocabulary learning and instruction including approaches to word knowledge and its dimensions, their representation in the mental lexicon, and shed light on the key concepts and features of formulaic language and collocations, as well as ESP vocabulary instruction.

We have seen that vocabulary learning cannot be regarded as an 'all-or-nothing phenomenon' (Melka, 1997, p. 87) as lexical knowledge is a continuum with different levels and dimensions in which some aspects of lexical awareness occur at productive, while others at receptive level. Lexical development is an ongoing process during which learners acquire not only word meanings, but make connections to other words, and learn how to use them appropriately in different contexts, including professional and academic settings. As collocational knowledge is an important component and significant marker of a proficient language user's language competence, the enhancement of learners' collocational awareness is a key aspect of successful language instruction. Additionally, the acquisition of collocations and formulaic patterns enhances students' motivation, as their overall proficiency improves in line with their collocational competence.

To answer the question of 'what' vocabulary to teach in professional and academic context seems to remain a never-ending challenge for ESP teachers. Since there is strong relationship between background knowledge and lexical knowledge (Marzano, 2004, pp. 1-16), learning profession-related words through content-based instruction is a highly effective way of vocabulary extension. Contextual ESP learning promotes that learners acquire professional communicative skills, and, become more autonomous and confident to take responsibility in their general and professional language development.

Furthermore, in the era of artificial intelligence (AI) that has had an influence on foreign language teaching methods, assessment, and the way of monitoring students' performance, the application of AI technology in vocabulary acquisition can be an effective complementary means of the traditional word-formation vocabulary teaching. The application of AI devices, building on and exploiting learners' existing knowledge, promotes cognitive development and promotes information processing ability (Lei, 2021, p. 659).

CORPUS LINGUISTICS AND CORPUS APPLICATIONS IN LANGUAGE TEACHING

3.1. Introduction

Corpus linguistics (CL) that analyses and exploits corpora, large systematic collections of spoken and written texts stored in electronic databases, is one of the most exciting and versatile approaches to contemporary studies and research in applied linguistics. Since the quantitative beginnings in the 1960s and the boom of corpus studies from the 1980s onwards, corpus linguistics has become a widely used research methodology in diverse applied linguistic domains, including lexicographic, lexical, grammatical, translation or media studies, text analysis, discourse analysis, semantics, pragmatics, sociolinguistics, and language pedagogy. Despite Chomsky's (1957) criticism of the emerging corpus-based approach to language studies that relies on the observation of naturally occurring data, corpus linguistics has become an independent domain of research and an influential approach to linguistic inquiry. Chomsky's rationalist paradigm made distinction between competence and performance, or internalized and externalized language, and considered the former the main objective of linguistic inquiry. In Chomsky's view performance or externalized language is affected by different factors that may impede competence, thus, he regarded corpora as collections of language performance that hinder the description of rationalistic cognitive models of performance (McEnery & Wilson, 1996, pp. 4-8). As CL follows a phraseological approach to analysis taking a syntagmatic view of language, Leech (1992) considers corpus linguistics a contrastive paradigm of Chomskyan generative linguistics. At the same time, CL is more than an innovative approach supporting and challenging empirical investigations of language variation and use, as it also enables researchers to examine features of naturally occurring texts from fundamentally different perspectives. To evaluate the information gained from corpus data, quantitative and qualitative research methods are extensively used in combination; it is characteristic of CL to begin with quantitative investigations of samples, followed by qualitative scrutiny of the findings that can lead to further quantitative analyses.

In the 1980s and 1990s, as McEnery and Wilson (2001, p. 24) pointed out, CL matured methodologically, and the range of languages and aspects of language use addressed by corpus linguists increased year by year (Biber, 1988, 1993; Flowerdew, 1992; Biber, Conrad & Reppen, 1998; Cobb, 1999; Conrad 1999). Consequently, researchers became interested in and fostered the application of the findings of corpus-based studies in language teaching in the early 1990s (Johns, 1991; Sinclair, 1991; Flowerdew, 1993; Widdowson, 1991), supporting the view that CL could revolutionize language teaching, syllabi and material design. Since then, the number of corpus-based studies and teaching materials has grown considerably. Research on the structures and use of language, as well as on the application of corpus-based findings covers a wide range of issues related to language pedagogy (Aston, 2000; Coxhead, 2000; Johns, 2002; Nesselhauf, 2004; Gabrielatos, 2005, 2018; O'Keefe, McCarthy and Carter, 2007; Römer, 2007; Aijmer, 2009; Henriksen, 2013). Practitioners of English for specific purposes were among the first ones who welcomed and appreciated the potentials of corpus-based language teaching in ESP classes (Flowerdew, 1993; Tribble, 1997; Gledhill, 2000). Corpus linguistics continues to gain acclaim with ESP practitioners, an increasing number of researchers explore the innovative ways of corpus-based methods in special language teaching (Bowker & Pearson, 2002, Coxhead, 2002, 2012; 2017; Gavioli, 2005; Marco, 2002; Kang, 2011; Flowerdew, 2004, 2012; Fujita, 2009; Hou, 2014; Fuentes, 2015; Jablonkai & Cebron, 2017).

The language of tourism is a multifaceted specific language with particular lexical, syntactic, pragmatic and stylistic features. In educational context where the primary aim is the acquisition of tourism discourse, monolingual specialised corpora can be exploited as authentic resources to explore different aspects of this special language. Employing parallel corpora how to teach the translation of tourism texts through the analysis of comparable concordances and collocations, as well as teaching corpus linguistics methodologies and procedures for the interpretation of data, are potential didactic applications of ESP corpora.

In this chapter I provide a brief overview of the development of corpus linguistics, discuss the main characteristics of the different types of corpora, shed light on distinctive features of specialised corpora, and present some key analytical corpus tools and methods. Then I move on to the discussion of the methodology of corpus linguistics, including the pros and cons of the corpus-based approach, the role of corpora in lexicography, and the key considerations of data analysis. The main aspects of corpus development are highlighted, and some guidelines are provided for the construction of specialised corpora. Finally, I discuss the relevance of corpora in language teaching, the pedagogical implications of corpora in vocabulary teaching,

focussing on vocabulary and text coverage, collocations, and the application of corpora in ESP classes.

3.2. What corpora are

3.2.1. The progress of corpora

Prior to the occurrence of electronic corpora, manual concordancing and indexing on paper slips required pain-staking work from literary and biblical scholars and lexicographers, furthermore, this time-consuming process did not allow the analysis of millions of words. The first comprehensive dictionary of English, a milestone work compiled by Dr Samuel Johnson, was published in 1775 as a result of several years' work with samples recorded on slips of paper. Probably the most famous dictionary based on a 'corpus of millions of slips of paper', amassed by the 1880s, is the *Oxford English Dictionary* that reflects the historical development of English language. At the time, the term corpus referred merely to a collection of written works similar in nature, or simply bound together (O'Keefe & McCarty, 2010, pp. 4-5).

Linguistics made headway in corpus-based descriptions in the nineteenth century with the studies of child language based on parental diary recordings, as well as in the early twentieth century with large-sample longitudinal studies of children's language acquisition. In the 1940s linguists began to use corpora in foreign language pedagogy; vocabulary lists for foreign language learners were often being derived from corpus data (Kennedy, 1992). It was in the 1950s when American structuralists came up with the idea of collecting real data for linguistic studies, thereby, the first computer-generated concordances stored with punch-card system occurred at the end of the decade. However, in the late 1950s corpora as sources of data underwent a period of unpopularity and almost total neglect, due to Chomsky's (1957) criticism of the corpus as a source of information, changing the directions of linguistics from the empiricism of corpus-based research towards rationalism. Yet, some linguists did not abandon using corpus-based methodology, in 1961 Quirk began to construct the most recent pre-electronic corpus, the *Survey of English Usage* (SEU), and in the same year Francis and Kučera started to work on the *Brown* corpus that took nearly two decades to complete. From this period the computer slowly started to become the main support of corpus linguistics, providing machine readable corpora for linguistic research, however, it was only in the 1980s and 1990s when corpora appeared as widely available tools for applied linguists (McEnery & Wilson,

1996, p. 4, 17). Since then, due to the technological advancements, the number and size of corpora has increased dramatically, and nowadays corpus methodology with its widespread popularity has opened up several new areas of research.

Concerning the history of corpus design, early or first-generation corpora that were developed in the 1950s, 1960s and 1970s are often distinguished from the larger late or second-generation corpora developed from 1980s onwards. First generation corpora include the first modern corpus, the *Brown Corpus* comprising American English texts, the *London-Lund Corpus of Spoken English* (LLC), and the *Lancaster-Oslo/Bergen Corpus* (LOB) based on written British English. For the first time, the parallel design of the *Brown corpus* and the *LOB* corpus made corpus-based comparisons possible between American and British English. Reflecting the technological constraints at the time, early corpora were generally limited to a one-million-word threshold (Adolphs & Lin, 2013, p. 598).

Second-generation corpora are either mega corpora comprising 100 million words or above, or monitor corpora. Two influential large-scale corpus projects were accomplished in the 1980s and 1990s, the *British National Corpus* (BNC), and the *Collins and Birmingham University International Language Database* (COBUILD), also referred to as the *Bank of English*. The COBUILD corpus was developed as a monitor corpus, this means that new texts are constantly added to its database, and the *BNC*, a 100-million-word corpus of modern British English contains samples of written language (90%) and transcribed spoken texts (10%). A significant mega-corpus, the *International Corpus of English* (ICE), initiated in the 1990s, consists of 22 one-million-word corpora respectively, representing regional varieties of English. Late monitor corpora that continuously expand in size capture the changes of language, as the *Corpus of American English* (COCA) and the *American National Corpus* (ANC). One of the main aims of these corpora is to provide textual database to the compilation of dictionaries and lexicographic research, as they focus on the use of general English (Adolphs & Lin, 2013, pp. 598-599).

3.2.2. What is a corpus?

The word *corpus* is of Latin origin, the *Oxford Latin Dictionary* defines it as 'any structure comparable to a body', or in more detail, 'a comprehensible collection of facts on a given subject: a compendium of scientific, literary or other writings'.

The term *corpus* in the context of modern linguistics refers to more than a collection of texts in McEnery and Wilson's view (1996, p. 21), as it has more specific connotations, namely; (1) *sampling and representativeness*, (2) *finite size*, (3) *machine readable form*, (4) and *standard reference for the language variety* that it represents.

Nesselhauf (2011, p. 2) defined corpora as systematic collections of naturally occurring texts, in both written and spoken language. In her interpretation 'systematic means that the structure and contents of the corpus follow certain extralinguistic (sampling) principles', and the information on the composition of the corpus is available to the researcher.

Francis (1982, p. 7), a pioneer in modern corpus linguistics emphasised the representativeness of a corpus in his definition; it is 'a collection of texts assumed to be representative of a given language, dialect, or other subset of a language'.

Leech (1997/a, p. 1) considered a corpus 'a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering'. Likewise Francis (1982)**,** he also stressed the importance of representativeness:

> It should be added that computer corpora are rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) *representative* of some language or text type (Leech, 1992, p. 116).

In Bowker and Pearson's interpretation (2002, p. 9) 'a corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria'.

Sinclair (2005) also highlighted the purpose and criteria of sampling in his definition:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (p. 46).

Flowerdew (2012/a, p. 3) epitomises the notion of a corpus with its characteristics; it contains authentic, naturally occurring data that are representative of a particular language and genre, and it is assembled on the grounds of explicit design criteria with specific linguistic or sociopragmatic purpose.

According to Pérez-Paredes's (2020, p. 11) present-time definition a corpus models language use and can be considered as a proxy for usage, an instrument and a method used to answer research questions.

Although there are several ways to define a corpus, McEnery et al. (2006) agreed that there is 'an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic texts* (including transcripts of spoken data), which is (3) *sampled* to be (4) *representative* of a particular language or language variety (p. 5)'. There have been arguments that subcorpora cannot be considered corpora in a real sense, nonetheless, McEnery et al. (2006) maintained that the term corpus, to the extent the collection of texts is useful and meet these four criteria, should always be regarded as a somewhat vague and inclusive term.

3.2.3. Types of corpora and their coded information

3.2.3.1. Diversity of corpora

As Horváth (2001, p. 38) claims 'the rationale of corpus linguistics is to directly access, derive, and manipulate evidence from a collection of texts'. These collections are compiled to serve different purposes, they differ in size and shape, and accordingly, can be classified based on various criteria. Corpora can be categorised on the basis of the source of the content, text types, number of languages, and inclusion of metadata, or its relation to other corpora. A corpus can fall into more than one category in case it complies with the criteria of those categories.

Researchers, lexicographers, experts and language teachers interested in applying corpus methods can rely either on general or specialized corpora. *General* corpora provide a text base for unspecified linguistic research, reflecting the use of a language or language variety in its contexts (e.g. *The American National Corpus*), while s*pecialized* corpora investigate some particular aspects of language, such as dialect, regional, or learner corpora, as the *Michigan Corpus of Academic Spoken English* (MICASE) and its British counterpart, the *British Academic Spoken English Corpus* (BASE), or the *JPU Corpus* of written learner English**,** to illustrate.

Concerning size, corpora that have a fixed size are referred to as *reference* corpora, as the *British National Corpus* (BNC), for instance. Unlike static corpora, the expandable dynamic corpora, also known as *monitor* corpora, are huge open-ended language 'banks' for linguistic

research that are constantly updated, like the *Corpus of Contemporary American English* (COCA). In contrast to monitor corpora, balanced corpora, also referred to as *sample* corpora represent a particular type of language over a span of time. Corpora that aim to achieve balance and representativeness in a sampling frame are called *snapshot* corpora, as the *Lancaster-Oslo/Bergen* (LOB) corpus (McEnery & Hardy, 2011, pp. 9-10). In terms of temporal dimension, *synchronic* corpora store samples exclusively from a given period of time, whereas the texts of *diachronic* (or historical) corpora illustrate the development of a language over time, as the texts of the same language are gathered from different time periods.

Additionally, corpora can be *monolingual* or *multilingual*. Monolingual is the most frequent type of corpus that is usually tagged for parts of speech, and it is exploited by a wide range of users. Multilingual corpora can be made up of same text-types in different languages, or can be *parallel* corpora, that is, two monolingual corpora that contain the translations of the same texts in different languages. A *comparable* corpus is made up of two or more monolingual corpora whose texts are related to the same topic, yet, they are not translations of each other. When users investigate these corpora, they can rely on the fact that they have the same metadata.

Other design-related distinctions can be made grounded on text types between corpora that contain *written* or *spoken* language, language produced by *native* or *non-native speakers* (generally learners), and corpora containing *whole texts* or *extracts with a specified length*. *Multimodal* corpora contain texts that are enhanced with audio or visual materials, or other types of multimedia content, adding further meaning to utterances and discourse of spoken corpora. For instance, the spoken part of the *British National Corpus* has links to corresponding recordings that can be played from the *Sketch Engine* interface (Adolphs & Lin, 2013; Horváth, 2001; Szirmai, 2005).

3.2.3.2. Corpus annotation and metadata

The utility and efficiency of a corpus can be increased by adding machine-readable coded information to the texts; tags or labels can be attached to the entire corpus, or to sections, sentences, phrases and words. Information related to the non-linguistic properties of texts is referred to as *mark-up*, whereas information on the linguistic peculiarities of texts is called *annotation*. Corpus annotation considerably extends the range of corpus-addressed research questions, therefore the information to be extracted determines what kind of mark-up or annotation is given to the corpus. Most corpora are annotated for *part of speech* (POS tagging) and different word forms, such as singular or plural, this kind of annotation by adding a

metalanguage is called *grammatical tagging*. Corpora can also be annotated from *lexical* aspect (e. g. lexis denoting assumption or prediction), or for *pragmatic* sense such as intensions or beliefs (Gabrielatos, 2005, p. 3). *Stylistic annotation* focuses on the stylistic features of literary texts, whereas *co-reference annotation* is a type of discourse level annotation. *Error tagging* assigns codes to types of errors that occur in a learner corpus, they are particularly useful when investigating features of non-native language behaviour (McEnery, Xiao & Tono, 2006, pp. 40-42).

Lemmatization, a type of annotation, is the process of reducing the inflectional variants of a word to its common base form (lemma). Lemmatization is important and beneficial in lexicography and vocabulary studies to improve dictionaries and computerised lexicon, yet, a lemmatizer ignores that some particular inflectional forms are used in particular collocations. However, as English language has simple inflectional morphology only a few corpora are lemmatized, despite the availability of various software to perform this process (McEnery, Xiao & Tono, 2006, pp. 35-36).

Besides assembling written and spoken language samples into a corpus, it is also important to document further information about the collected texts. Metadata, or 'data about data' provide editorial, analytic, descriptive and administrative information about the corpus itself. *Editorial metadata* provide information about the relationship of corpus components and their original source, while *analytic metadata* inform about the ways of interpreting corpus components. *Descriptive metadata* provide classificatory information of corpus components, whereas *administrative metadata* provide documentary information about the corpus, such as its availability, or its revision status. Metadata are crucial to meet the standards of *representativeness*, *balance* and *homogeneity*, and they are particularly important in sharing and reusing a corpus, since the documentation and design rationale facilitate the replicability of the research and the validation of results (Adolphs & Lin, 2013, p. 599).

3.2.4. Specialized corpora

3.2.4.1. Types of specialized corpora

ESP settings are one of the most obvious applications of corpora and corpus methods. Specialized corpora contain texts of a certain type as they aim to be representative of a specific language highlighting how that specific language works, and are assembled to answer specific

questions. Since they are domain or genre specific and designed to represent a sub-language, they provide valuable resources for investigations in the relevant fields and genres, and they can be created from a subcorpus of general corpora that contains texts limited to a subject area. Specialized corpora can be distinguished as *language for special purposes corpora* (LSP), *learner corpora*, and *parallel or translation corpora* (Römer, 2007, p. 6).

A learner corpus is directly related to the language classroom as it is a collection of writings or speech of learners who acquire a second language (L2). Learner corpora capture the language use of second language learners and the developmental aspects of learner language, thus, they are prone to be erroneous. Learner corpora have been used to fulfil two distinct yet related functions; they provide a better understanding of interlanguage and this way contribute to SLA research, and highlight those factors in learners' speech and writing that influence interlanguage development. This way, learner corpora can considerably contribute to the development of pedagogical tools and methods adjusted to the needs of language learners (Granger, 2008, p. 259).

Studies on learner corpora highlight the specific problems of language learners, and help teachers to pinpoint whether a particular language phenomenon is difficult or not (Römer, 2007, p. 7). According to Flowerdew (2001, pp. 376-77) they 'can provide useful insights into which collocational, pragmatic or discourse features should be addressed in material design'. The ICLE, *International Corpus of Learner English*, the BLC, *Learner Business Letter Corpus*, and the *JPU Corpus* of written learner English set example to the succeeding compilers of learner corpora (Nesselhauf, 2004, p. 151). The MICUSP, *Michigan Corpus of Upper- level Student Papers*, assembles texts from a range of university disciplines. Corpora that provide basis for investigating the use of English as lingua franca include the *English as a Lingua Franca in Academic Settings* (ELFA) corpus, and the *Vienna - Oxford International Corpus of English* (VOICE) (Adolphs & Lin, 2013, p. 599). *Developmental* corpora track the changes in language use of children acquiring their first language (L1); a well-known example is the *Child Language Data Exchange System* (CHILDES) corpus.

Parallel corpora provide basis for studying how meaning in one language is conveyed to another language. Contrastive analyses of parallel or translation corpora are invaluable for the selection of the different and mistreating elements of learners' native language and their target language equivalents, additionally; they are important tools in the training of interpreters and translators. Another application of parallel corpora is the investigation of collocational and

phraseological differences of the target and native languages (Römer, 2007, pp. 6-11). In order to identify which sentences or words are translations of each other, parallel subcorpora need to be aligned, this alignment can be achieved using statistical measures and mutual information scores (McEnery & Wilson, 1996, p. 58). To illustrate, McEnery et al. used an English-Chinese parallel corpus to explore how the expressions of temporal and aspectual meanings differ in Chinese (McEnery, Xiao & Tono, 2006, p. 91).

LSP corpora have a great impact on syllabus and material design, and can help solving the problem of selecting the appropriate vocabulary for ESP courses; as Gavioli (2006, p. 23) remarked 'working out basic items to be dealt with is a key teaching problem'. Since LSP corpora comprise texts written by experts of the subject field, learners can delve into a body of evidence related to the function and usage of lexical items in that domain. In addition, with the help of corpus analysis tools meaningful patterns can be revealed in different context, and learners can also observe how a term or expression *cannot* be used (Bowker & Pearson, 2002, p.19). Moreover, frequency and concordance data of a corpus enable teachers to select those words and expressions that learners need in order to handle texts of their subject area. Well-known examples of the exploitation of LSP corpora are the *Academic World List* (AWL) comprising the core vocabulary used in academic texts, compiled by Coxhead in 2000 (Bennett, 2010, pp. 11-13), the *Academic Formulas List* (Simpson-Vlach & Ellis, 2010), and more recently the new *Academic Vocabulary List* (AVL) derived from the academic subcorpus of the 425-million-word *Corpus of Contemporary American English* (COCA) (Gardner & Davies, 2014, p. 305).

3.2.4.2. Tourism corpora

As Dann's (1996) investigations on the specific linguistic features of tourism English were not based on corpus evidence, Castello (2002) and Lam (2004) can be considered the initiators of tourism texts analysis using a corpus-based approach. Lam (2004) compiled a tourism corpus of written texts; the *Tourism Industry Text Corpus* (TITC) comprises 2 million tokens, unfortunately not a public domain, that covers the main registers of tourism texts including travel guides, tourist information, travelogues, and promotional texts. He conducted corpus-driven research on the lexico-grammatical features of the language of tourism; the lexical keywords were related to different semantic fields, then the collocates of each target word were categorized semantically, and additional prosodic meanings were observed in the concordance lines. His analysis revealed differences between general English and English in tourism

language from an educational perspective, providing some important pedagogical implications in ESP and lexicography.

The innovative *NESPOLE! VoIP Multilingual Corpora in Tourism and Medical domains* were designed to provide human-to-human dialogues in real word settings. The multilingual VoIP (Voice over Internet Protocol networks) database consists of 200 recorded and transcribed dialogues in four languages (English, German, French and Italian). The tourism subcorpus comprises 66 monolingual dialogues between a travel agent and a prospective tourist. These speech corpora represent valuable resources for phonological and linguistic research (Mana et. al, 2002).

Besides the praiseworthy compilation of several small-scale corpora to investigate different features of tourism language, there have been a few remarkable undertakings in the development of parallel tourism corpora and corpus-based comparative studies of tourism language. Wilkinson (2006) examined the utilization of *parallel English-Finnish tourism* corpora and their integration to the training of translators. His study drew attention to the fact that electronic corpora can be powerful performance enhancing resources and help students to confirm or reject decisions based on other tools, including dictionaries. The monolingual *TourEC* (Tourism English Corpus) was compiled in 2011 in Italy with focus on authentic communication in tourism; it comprises travel articles written in English on a variety of topics such as tourist cities, countries and geographical regions. The multilingual *T-TourEC* (Translational Tourism English Corpus) was created in 2012 within the same project; it is made up of texts translated into English from a variety of languages including Italian, Norwegian and Japanese. The *English-Hindi Parallel Tourism Text Corpus* contains descriptions and information pertaining to tourism destinations and tourist interests. A more recent enterprise is the building of an *English-Cebuano (Philippine) parallel tourism corpus* (Bureros, Tabaranza & Roxas, 2017) based on bilingual websites. Related to the main concepts of tourism such as destinations, catering, or festivals, a separate named entity list, that is, a list of names with bilingual entity tags was generated from the corpus.

3. 3. Corpus linguistics, corpus tools, methods and techniques

3.3.1. Corpus linguistics approach

The initiatives of corpus linguistics are nearly contemporaneous with Firth's (1957) collocation theory, since then the advancements of information technology enabled empirical investigations of significantly larger dataset, eliminating the inadequacies of mere intuition in the identification of grammatical structures or lexical sets. Concerning English language, its globalization and recent changes cannot be detected and traced solely by intuitions of researchers (Krishnamurthy, 2001, p. 2). As Hunston and Francis (2000, p. 14) argued 'corpus linguistics… prioritizes a method, or groups of methods, and a kind of data, rather than theory'. Owing to the primacy of methods in which discoveries precede theories, the bottom-up approach is an inherent feature of CL methodology, opposed to top-down methodology where theory designates the scope and terms of research (Krishnamurthy, 2001, p. 2).

Corpus linguistics as a methodology of linguistic analysis considers naturally occurring language samples a reliable source for linguistic research, and can be used for the investigation of numerous kinds of linguistic queries (Nesselhauf, 2011, p. 2). Opposed to this view, Chomsky (interviewed by Andor, 2004) rejected the credibility of observed data as evidence to support linguistic claims; 'Corpus linguistics doesn't mean anything' (p. 97). In Widdowson's view (2003) applied linguistics, and within its framework corpus linguistics, can function as a mediator to connect theoretical linguistics and language teaching practices. To fulfil this function, CL as a system of methods and principles how to apply corpora in linguistic inquiry, puts into pedagogical practice the implications of corpus-based research.

However, as McCarthy and O'Keeffe (2010) remarked, corpus linguistics is associated by some linguists as a 'method of exegesis based on detailed searches for words and phrases in multiple contexts across large amounts of text' (p. 3). Nevertheless, corpus linguistics has provided a wealth of empirical data available to language teachers; the information about frequency, syntactic and lexical patterns can play an important role in deciding what is more or less important in a language classroom (Keck, 2013, p. 1). Language testing and teacher education are other areas within applied linguistics where there has been a rapid growth in the application and development corpora.

3.3.2. Corpus tools and methods

Numerous multi-faceted user-friendly software packages are available for the extraction, analysis and manipulation of corpus data. The most frequently used functionalities include the generation of frequency counts on the basis of different criteria, comparisons of frequency information in different texts, concordance outputs in different formats, including the *Key Word in Context* (KWIC) concordance, and the extraction of word clusters in a text. Several programs can be downloaded from the internet or can be used online, such as the *Compleat Lexical Tutor*, *Antconc*, *BNCWeb*, *SketchEngine*, or *Wordsmith* (Adolphs & Lin, 2013).

The frequency of words and multi-word units in different contexts provides important information not only on their occurrence, but different word forms of the same lemma. Word lists, which are based to certain extent on frequency, are good baselines for subsequent searches of individual vocabulary items at concordance level, and they are adequate for the comparison of different corpora. Lemmatized frequency lists assemble inflectional and derivational forms of the same lemma; the preceding lemmatization can be done manually based on an alphabetical frequency list, or in automated way using lists of preselected lemmas. These lemmatised lists can be applied effectively in foreign language teaching, it is beneficial to teach all variations of a lemma together focussing on the most frequently used forms (Adolphs & Lin, 2013, p. 600).

Keywords are those words that occur either with a significantly higher or significantly lower frequency in a text or texts compared to a reference corpus; they are derived from the statistical comparison of the target or reference corpus; the statistical significance of difference is calculated with *chi-square* or *log-likelihood* statistics. Both statistical measures compare the actual observed frequencies with the expected frequencies of the items, assuming random distribution; however, in case the difference between them is large it is likely that this relationship is not random. Thus, chi-square and log-likelihood statistics generate words that are characteristic or uncharacteristic in a particular corpus, subsequently, keyword lists and lists of key sequences can be generated based on the identified items (Adolphs & Lin, 2013, pp. 601-602).

Presumably the most widely used corpus tool is the *Key Word in Context* (KWIC) concordancer that provides an alphabetical listing of the key-word called node or target item and its collocates. The concordance output displays the specified number of words to the left and to the right of the node word, each occurrence of the target word or phrase is centered in the middle,

surrounded with the relevant part of co-text. A short space, usually the distance of no more than four words to the right and left of the node, is referred to as span (Sinclair, 1991, p.170).

Figure 1: *Screenshot of an Antconc 3.2.4w Concordance Window on collocates of '**building**'*

If a corpus includes part-of-speech (POS) tagging, users can search for items or patterns based on word class or grammatical structures. Concordance outputs provide a representation of language data, allowing the users to examine patterns in which a lexical item or a cluster is used in real language context (Adolphs & Lin, 2013, p. 602).

3.4. Corpus development

3.4.1. Guiding principles of corpus design

Various factors need to be considered before the actual corpus compilation begins. The fundamental question to address before commencing the compilation of a corpus is as to whether it is necessary to do so; 'So before tackling the task of building a corpus, be sure that there is not an existing corpus that meets your needs (Reppen, 2010, p. 31). Since a corpus is built for the study of language, a well-designed corpus will represent and support this purpose. When collecting data, we need to take into consideration what amount of data and what methodologies allow us to answer particular research questions; what kind of query is anticipated, and what kind of methodology to use to study the data. The compiler needs to consider and determine the size of the corpus, the text types that should be included, the number

of samples for each text type, and the length of the samples. Once the decisions are made, it should be taken into account how a corpus can be most effectively created, how to collect texts for inclusion, keep records of the collected texts, obtain permission for copyrighted texts, and store them in a selected type of electronic format (Meyer, 2004).

Corpus design and the creation of the corpus are matters of paramount importance, since the findings of corpus-based studies can be interpreted only on the grounds of the examined texts. Influential theoretical and practical considerations of corpus compilation have been suggested in CL literature (Biber, 1993; Biber et al., 1998; Leech, 1997; Sinclair, 2001, 2005; Szirmai, 2005). One of the most determinative set of guidelines was proposed by Sinclair (2005) who defined ten key principles for developing corpora:

1. The contents of corpora should be selected according to their communicative function in a particular community in which they occur.
2. Corpus should be as representative as possible.
3. Only those components of corpora should be contrasted that have been designed to be contrastive.
4. Criteria for determining the structure of a corpus should be small in number… delineating a corpus that is representative of the language or variety under examination.
5. Any information about the texts apart from the alphabetical listing of its words should be stored separately from the main texts and intermingled when required in applications.
6. Samples of language should preferably consist of entire documents or transcripts of speech events; this means that samples will differ substantially in size.
7. The design and composition of a corpus should be documented fully with information about the content with arguments on the decisions taken.
8. The corpus builder should hold on to representativeness and balance, they must be used as guidelines of the design of the corpus and the selection of its components.
9. Any control of subject matter in a corpus should be based on external, not internal criteria.
10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and useless texts should be avoided. (adapted from Sinclair, 2005, pp. 1-25)

Such design criteria are not only applicable, but ensure that the data will be valid and reliable for analysis and interpretation by the corpus researcher, 'the results are only as good as the corpus' (Sinclair, 1991, p. 9). Biber et al. (1998) advocate to be realistic, careful, and flexible in corpus design:

> Given constraints on time, finances and availability of texts, compromises often have to be made. Every corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues (p.250).

Clear (1992, p. 30) advocates that decisions concerning the quality of the corpus possibly be based on assessment of existing corpus resources, and corpus creation should be based on experience gained from previous corpora in terms of linguistic description and methodology.

3.4.2. Size, representativeness, and balance

To make a decision concerning the size of the corpus is a difficult task. Grounded on the interrelation of research question, representativeness and design, the question of size can be resolved by taking into account two factors, *representativeness* and *practicality*. To achieve representativeness, the collected texts need to represent exactly the type of language under scrutiny, yet the time constraints delimit the size of the corpus. However, in most cases it is impossible to achieve complete representation, this way the size of the corpus is determined by collecting a sufficient number of samples of the language for intensive representation (Reppen, 2010, pp. 31-32). On the other hand, the eventual size of the corpus depends on, to a great extent, the number of available resources, as Nelson (2010) argues 'any attempt of corpus creation is therefore a compromise between the hoped for and the achievable' (p. 60). Sinclair (2004) advises that it is important to avoid perfectionism when compiling a corpus, as corpus linguistics is an inexact science no-one knows what exactly an ideal corpus should be like.

Since the objective of corpus research is to make generalizations about the language, representativeness and balance are the basic requirements of a corpus (Szirmai, 2001, p. 7). Accordingly, corpus builders aim to accomplish the requirements of balance, representativeness and comparability of their corpora, that, in fact, they rarely attain. Leech (1991, p. 27) considered a corpus representative when 'the findings based on its contents can be generalized to a larger hypothetical corpus'; though he remarked that the assumption of representativeness 'must be regarded largely as an act of faith'. In Szirmai's (2001, p. 12) view the size of the corpus is in close correlation with representativeness and with a principled sampling procedure; when the texts are assembled from the same author and genre, or share other characteristics, maximal representativeness can be easily achieved. However, size itself does not guarantee representativeness, corpora should be 'as large as necessary and as small as possible' (Biber, 1993, pp. 242-246). Leech (1997/b, p.22) also holds the stance that large size is not always advantageous, and smaller corpora can prove to be very effective if accurately selected. Biber (1993) maintains that representativeness can be realized if the corpus entirely captures the variability of language. As Sinclair (2005) pointed out the degree of specialization is also a

significant factor as 'a specialized text in a general corpus can give the impression of imbalance' (p. 40).

3.4.3. Steps of corpus design and compilation

As discussed in 3.4.1., descriptions, theoretical and practical advice for corpus development have been provided in the literature of corpus research since the 1990s (Horváth, 2001, 2005; Kennedy, 1998; Sinclair, 2005; Szirmai, 2005). In Sinclair's (2005) view the two stages of corpus creation, *design* and *implementation*, are barely separable, nevertheless, it is important to consider soundly the individual steps and process of corpus design, and its implementation. These steps require theoretical and practical considerations as well, include different procedures and several decisions to make, in order to achieve the ultimate purpose of the corpus compilation and analysis.

According to Horváth (2001) a corpus can be set up on the basis of well-defined factors of design such as 'purpose, language community, text types, representativeness, encoding and storage facilities' (p.101). Horváth (2006, pp. 112-114) provided useful guidelines for corpus development and summarized the stages of corpus compilation as follows; 'On the basis of the billions and billions of words spoken and written down, corpus linguists do the five Ss: they *select*, *structure*, *store*, *sort* and *scrutinize* language (p. 112)'. After defining the reason of corpus development and its objectives have been set up, the collection of materials can commence. During the phase of structuring the corpus has to be categorized by different aspects, and the success of the forthcoming research, to a great extent, depends on the proper saving and storing of texts. The first four steps provide the basis of the last phase, scrutinizing, that is, the examination of the corpus on specified queries.

To follow *Biber's model of cyclical corpus* design was proposed by Horváth (2001, p. 45) in order to compile a valid and reliable corpus. The development of a sampling frame ensures that we be able to use data of the represented population that the research intends to study. As *Figure 2* illustrates, first a pilot study should be conducted to gain information on the effort of the population before the next step, corpus design proper.

```
pilot empirical                Corpus          compile        empirical
investigation and  →  design  →  portion of  →  investigation
theoretical analysis           ↑               corpus
                               └───────────────────────┘
```
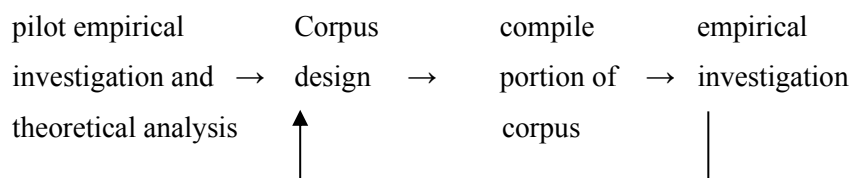
Figure 2: Biber's adapted model of cyclical corpus design by Horváth (2001, p. 45)

Corpus design proper involves the specification of the length of each component in the text, the range of text types, the number of individual texts, and testing of a random selection technique in order to provide equal chance to each potential texts to be selected for the corpus. The third stage is the assembly of a subcorpus in which the specifications are tested. During the final procedure, empirical investigation is carried out on the reliability and representativeness of the corpus, the specifications studied are compared with the samples, and statistical measurements are administered.

3.4.4. Design of specialised corpora

The first important step when creating a specialized corpus is to take into consideration the teaching context, the students' language needs, and how the corpus will be exploited and used. Flowerdew (2004, pp. 25-26) presented some general guidelines for the construction of specialised corpora and some key considerations that need to be taken into account. Specialised corpora are built with a preconceived purpose in mind that determines, to a large extent, what genres are to be investigated. The appropriate size of a specialised corpus highly depends on the investigated phenomenon; the lower the frequency of the feature under scrutiny the larger the corpus, and conversely, when one intends to investigate more common grammatical items the corpus can be smaller. A pilot survey should be conducted to reveal as to whether the corpus yields sufficient amount of data for the analysis of the examined patterns.

As in the case of designing general corpora, the most important consideration in planning a small specialized corpus is representativeness. Biber (1993, p. 243 quoted in Koester, 2010, p. 69) specified representativeness as 'the extent to which a sample includes the full range of variability in a population', and he made distinction between *situational* and *linguistic* variability. Whereas situational variability refers to the registers and genres of the target population to be included in the corpus, linguistic variability implies the range of linguistic distributions in the population. The collected samples should meet both criteria with the

primacy of situational criterion, as a corpus cannot be linguistically representative without having established situational representativeness. Therefore, it is advisable to define the sampling frame, the entire population of the texts to be investigated. Obviously, there are limitations of sampling for a small corpus, as it will be impossible to collect samples that entirely cover a widespread genre. Yet, it is important to ensure that the included texts be typical in that particular domain and setting. Therefore, attentive sampling is essential to achieve representativeness, yet, it should be borne in mind that representativeness cannot be evaluated objectively, thus, when the results are skewed in some way it turns out that the corpus is *not* representative (Koester, 2010, p. 69).

When building a specialized corpus, it is reasonable to suppose that a corpus related to a certain subject area will have a greater concentration of specialized vocabulary than a broad-range general corpus. However, the number of different word forms is far less in specialized corpora than in general ones, and the proportion of single occurrences even shows a greater difference. Word forms that occur several times are much less common in a specialized corpus, suggesting that the specialized corpus unfolds smaller but technical vocabulary. Although specialized corpora are likely to contain fewer words, the characteristic vocabulary of the specific domain is prominently displayed in the frequency lists, indicating that a smaller corpus is sufficient for studying the features of specific languages (Sinclair, 2004).

Jablonkai (2010) proposed a *Model for corpus creation in ESP* and divided corpus design and implementation to seven stages that cover both theoretical and practical considerations. The ultimate purpose of the corpus determines the subsequent decisions concerning the steps and aspects of corpus design, such as the type of the corpus, or the text categories. Her model provides some guidelines concerning the size, representativeness, sampling, data collection and data entry. The suggested steps of corpus design are presented in *Figure 3*:
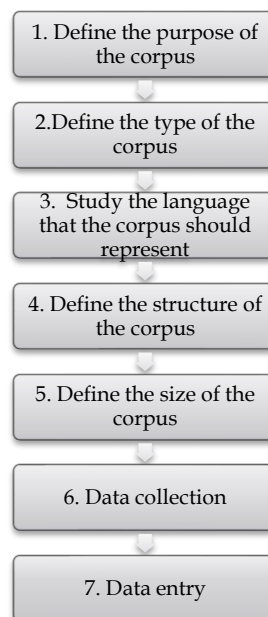
Figure 3: Model for corpus creation in ESP (Jablonkai, 2010, p.118)

The first three steps provide criteria to the selection of the texts; as a first step the purpose of the corpus should be identified, in Step 2 decision has to be made on the type of the corpus, and Step 3 involves studying the language that it should represent. The next two steps, decisions on the structure and size of the corpus are based on the previous decisions; however, they are governed by practical constraints. Step 6, data collection may require changes to the originally planned content in the proportion and in the inclusion or exclusion of certain text types. The final step, data entry involves decisions on entering the data into a database and creating a storage system.

3.5. Corpus linguistics in language pedagogy

3.5.1. Corpus-based and corpus-driven approaches to language pedagogy

Corpus linguistics is generally related to a phraseological approach of analysis, taking a syntagmatic view of language. In fact, corpus analysis allows both syntagmatic and paradigmatic view of language, as concordance outputs can be interpreted either paradigmatically, that is, lexical items can be substituted for one another, or syntagmatically, from a phraseological aspect. This phraseological approach is regarded as a *corpus-driven approach*, in which the corpus is used as evidence, and the analysis of corpus data is carried out without any pre-conceived notions or assumptions concerning how they should be analysed

(Flowerdew, 2009, p. 393). Opposed to the corpus-driven approach, in *corpus-based research* a theoretical statement precedes corpus investigations, and the primary aim is the analysis of pre-defined linguistic features of syntactic patterns and structures (Biber, 2012). In corpus-based approach the corpus is used to exemplify and quantify theories guided by intuition.

The application of corpus-based research in teaching English as a second language covers three main areas such as *indirect use* of corpora, *direct use* of corpora, and *compilation of corpora for educational purposes*. Corpora are being indirectly used for the development of corpus-based materials including word lists, dictionaries, grammar reference books, textbooks or lexical syllabi with focus on communicative competence (Tono, 2011, p. 3). Other *indirect applications of corpora* include language for specific purposes (LSP) corpora, learner, and translation corpora with different implications in the language classroom. ESP students can benefit from genre-specific corpus-driven glossaries or from authentic teaching materials derived from concordance data. Flowerdew (1993) suggested that the lists of most frequently occurring words of an ESP corpus should be used for selecting and grading items to include in a lexical syllabus. Learner and translation corpora offer a variety of practical uses including learner dictionaries or materials based on error analysis. The *direct use of corpora* involves the classroom application of corpora with the typical data-driven learning (DDL) approach; students are faced with hands-on tools like on-line corpora, or retrieve and discuss concordance lines related to a particular task (Casas-Pedrosa et. al, 2013, p. 2). The *compilation of corpora* for pedagogical purposes concerns different specific areas such as textbook corpora, learner corpora or corpora containing texts with controlled vocabulary (Tono, 2011, p. 3).

Corpus linguistics has provided the means of describing language in use in a more informative way, and has a real impact on the enhancement of pedagogical applications grounded on those descriptions. Despite critical arguments of using corpora in language teaching such as the isolated descriptions of language use, or the ignorance of the conceptual aspects of the entire text (Widdowson, 2000), corpus-based instruction undoubtedly facilitates access to naturally occurring language, providing information about real language use, and about the frequency and collocations of lexical items.

Although frequency is a crucial criterion of the selection of lexical items, the mere frequency of a language pattern does not necessarily justify that it should be prioritized in a language teaching syllabus (Widdowson, 1991). Granger (2011, pp. 131-132) claims that frequency

needs to be counter-balanced by other factors such as learnability, teachability and learner variables including age, aptitude, motivation or learning style. Cook (1998, p. 61) warned that giving prominence to frequently occurring words and expressions at the expense of rarer but more effective or salient ones might impoverish language learning. It is also contested whether it is beneficial to face learners with authentic language in use, in some cases real language might not be interesting and particularly useful for language learners; text types such as advertisements, leaflets or reader's letters are real only to the members of the target speech community (Carter & McCarthy, 1996).

Two main areas of language teaching and learning can benefit from corpora; incorporating the latest findings of corpus-based research in syllabi, teaching materials, dictionaries and language testing, and in the methodology of classroom management by encouraging teachers and language learners to examine language patterns of corpora in and outside classrooms. Furthermore, the application of corpus methods and techniques promotes exploratory, discovery and autonomous language learning. However, as Widdowson (1991) maintains it is not always expedient to utilize corpus data directly in pedagogic materials due to their cultural or contextual inappropriacy. He advocated to adapt some kind of 'pedagogic processing' to transform language samples into pedagogically applicable examples, and emphasised the need for some kind of 'mediating process' that facilitates adjusting corpus data to learners' socio-cultural and linguistic environment. In order to integrate this pedagogic processing in language learning, Flowerdew (2009) proposed peer activities in which more proficient students grouped with weaker students share their interpretation of corpus data, and assist the less proficient ones with collaborative dialogues to gradually develop their independence in language use.

3.5.2. Data-driven language teaching and learning

The use of corpora in language teaching and learning is generally associated with Tim Johns (1991). His ground-breaking work on data-driven learning (DDL) demonstrated and accentuated the relevance of discovering language with corpus analysis techniques, he suggested that teachers should 'confront the learner as directly as possible with corpus data, and make the learner a linguistic researcher' (2002, p. 108). Accordingly, DDL can be defined as a learner-centered approach to language learning in which students explore the authentic texts of a corpus making use of effective corpus tools, such as concordancers or keyword extractors. Johns (1991) introduced an interactive data-driven learning technique, the Kibbitzer

approach, to help students identifying mistakes in lexical choices. He provided parallel concordance lines using authentic and unmodified texts of large corpora; with the concordance patterns learners would discover themselves the alternative syntactic or lexical choices of the problematic items in various contexts (Dudley-Evans, 1998, p. 9).

The introduction of corpus data in the classroom has brought several challenges to language teachers; they need a lot of preparation, as well as they need to develop students' skills of investigation, then, since the investigations lead to observations and questions, teachers need to hand over control and promote implicit language learning (O'Keefe & McCarthy, 2010, p. 6). As DDL work generally places considerable demands on teachers, it is more suitable for advanced and motivated learners (Boulton, 2016, p. 7). In Johns's classes students were involved in accessing and exploring corpus data in individual or pair work activities by scrutinizing on-line concordances. Through the analysis of the authentic texts of a corpus, students can describe different features of a language, as well as they can investigate significant or problematic patterns in natural context.

According to Horváth (2001, p. 58), the DDL method with its 'remedial grammar' tools to correct usage of English may be regarded as a subdivision of CALL, computer-assisted language learning. Nevertheless, as Gabrielatos (2005, p. 9) remarked any exploration of the behaviour of lexical or grammatical items can be referred to as data-driven approach, this way distinction should be made among *text-based* DDL, *corpus-derived* and *corpus-based* DDL. Corpora can be introduced in the classroom either in the soft or hard version of corpus-derived and corpus-based data-driven learning (Leech, 1997/b, p. 10). Applying the soft version, it is the teacher who accesses the corpora, compiles the tasks and prepares printouts, whereas in the hard version learners have direct access to corpora that helps to adjust the focus of examination to students' needs and interests.

McEnery and Wilson (1996, p. 104) highlighted that foreign language teachers often expose students to simplified examples that will cause difficulties with processing when they are confronted with more complex language use. Corpus linguistics, for this reason, can contribute to more effective language learning as students can explore authentic texts and face with real-life language use. Additionally, CL provides real textual data to assess hypothesis about language use, allowing the exploration and investigation of linguistic structures that we aim to inquire into. The benefits of corpus-based vocabulary teaching and learning have been proven

by numerous studies (Cobb,1997; Creswell, 2007; Fuentes, 2004, 2007) highlighting the role of concordances in vocabulary extension and error correction.

Notwithstanding the advantages of the DDL approach, according to Flowerdew (2009, p. 395), there are some drawbacks of using corpora in DDL. On one hand, corpus linguistics techniques favour more a bottom-up rather than top-down processing, this way truncated concordance lines are examined in isolation. On the other hand, corpus data are decontextualized, therefore, they cannot be implemented directly in the context of students' own writings. Corpus-based learning is generally considered to be an inductive approach to learning in which rules are constructed on the basis of multiple examples, yet, this approach might not be appropriate for field-independent students who rely on rule-governed deductive approach. In addition, since there is an abundance of different types of corpora, language learners might have difficulties in selecting the most suitable resource for a specific query. Despite the many benefits of corpus linguistics students also need to be aware of its limitations; likewise other approaches and methods, corpora cannot be representative of all language, and cannot provide negative evidence of language use (Bennet, 2010).

3.5.3. How can a corpus help special language acquisition?

3.5.3.1. The role of corpora in ESP

Language learning can be regarded as an inductive process in which learners associate meanings to forms. This view corresponds with the *schema* theory of cognitive psychology on memory; a schema is an imprint of an event that is individualised and selected for remembering according to our interest and ideas. This way language learning is a process that involves matching real world knowledge structures to the appropriate and relevant pragmatic and linguistic schemata of the target language. As corpora provide access to authentic oral and written interactions, they allow learners to observe what is typically said in a situation and how it is typically expressed. The schema-meaning pairs are constructed on the grounds of repeated experiences of instances of the target language use, and with the help of a concordancer this process can be short-cut (Bernardini, 1996, p. 18). ESP learners need to draw on appropriate schematic knowledge, which is represented in their lexical choices, in order to interact efficiently in written and spoken academic or profession-related communication.

Corpora as teaching aids, learning tools and reference resources are particularly relevant to ESP (Bennet, 2010, p. 11). Bowker and Pearson (2002, p. 48) shared the view that even a small-scale corpus is a reliable aid when working with domain specific language, and in terms of application Biber (2006) provided some masterly examples how specialized corpora can be utilised, and how corpus-based investigations can arouse students' language learning interest. Due to the specificity of ESP, ready-made traditional resources such as course books and dictionaries are unlikely to cover the highly specific language learning needs of ESP users, or trace the changes of terminology in specific areas. A corpus comprising specialized texts can be used rapidly to generate a series of items and tasks characteristic to that particular text type; these items can provide basis for the selection of language features to be included in the ESP syllabus (Gavioli, 2005, p. 23). The analyses of the language of specific domains generally focus on the lexis; the investigations of the texts of specialised corpora result in word lists, keyword identification and collocation analysis.

Corpus linguistics gives insight into the conventionalised language used in specific disciplines and domains, and into their specific genres and text types. Since corpus tools and techniques have become increasingly available and user-friendly, teachers and learners can use them to uncover regularities at various levels, from discourse to collocations, to lexical bundles or special terminology (Boulton, 2012, pp. 261-262). Yet, as Boulton (2012, p. 275) concluded, corpora and DDL are not particularly effective for learning large amounts of words, they are more appropriate for refining usage in context, and enhance the depth rather than the breadth of vocabulary knowledge. This makes the corpus-based approach particularly useful for ESP students, who often have extensive referential knowledge of technical words related to their academic or professional field, but have difficulties in using them in context.

On the other hand, concerning special vocabulary, Durrant (2014) accentuates that corpora cannot provide us information about 'how students interact with the texts or what they need to be able to know about or do with words to complete their tasks successfully (p. 354)'. Due to this scantiness, qualitative corpus-based studies can shed light on the different aspects of learning specialized vocabulary. Fuentes (2004) created a 'professional' corpus of 25,000 words from spoken business reports, and an equivalent 'academic' corpus containing discussions on socioeconomics from the *MICASE*. The corpora were explored by tourism students to prepare for an individual oral report and to take part in a group discussion. The members of the experimental group used the target items from the corpus effectively in the

individual report, and they were more confident than the control group in the spontaneous discussion. Yet, the experimental group made more errors that can be attributed to their faster and longer presentations. Tourism students were also the participants of another study conducted by Fuentes (2007); the students spent three hours exploring register, based on six tourism advertisements, and afterwards, for two hours they examined concordances of other 60 adverts. The tasks were related mainly to register, text type, and vocabulary. The experimental group performed significantly better in the reading comprehension post-test in all task types than the control group, who used text books and texts in a traditional way.

3.5.3.2. Corpora in teaching the language of tourism

Since in teaching the language of tourism the acquisition of the special tourism discourse is the primary objective, specialised corpora can be valid and valuable resources to explore different aspects and typical characteristics of this specific language. Applying corpus processing software, teachers and students can explore and analyse a large amount of textual data as authentic references in language classes, this way acquire the specific linguistic features that a prospective tourism professional should be able to master. To teach the language of tourism at intermediate or advanced level (B2/C1) Gandin (2016, pp. 68-69) proposed three specific didactic scenarios:

1. *Teaching English tourism discourse*, its typical features and characteristics based on authentic materials coming from monolingual specialized corpora, so that learners can explore different aspects of tourism discourse including linguistic, stylistic, and pragmatic properties (e.g., specific verbal techniques or communicative functions).
2. *Teaching how to translate tourism texts;* with the help of parallel or monolingual translational corpora learners can identify the key translational features and strategies used in tourism texts, and compare specific lexis and collocations.
3. *Teaching corpus linguistics methodologies* by using tourism corpora to make learners familiar with corpus linguistics methodologies, corpus tools and procedures for the interpretation of data.

Corpus-based materials used as pedagogical resources allow language learners to observe the main properties of tourism language in context, and with the help of concordances they can detect the collocates of specific key-lemmas related to important subsets of tourism and semantic areas of tourism discourse, explore the phraseology of tourism, the lexico-grammatical choices and recurring patterns of tourism language. While analysing the main

properties of English tourism discourse through corpus-based examples, learners become familiar with corpus linguistics research tools and methodologies (Gandin, 2016, pp. 68-69).

3.6. Corpora in lexicography, phraseology and vocabulary selection

3.6.1. The role of corpora in lexicography and phraseology

Within the number of ways corpus linguistics has contributed to lexicography, probably the lexicographical treatment of phraseology has had the most revolutionary effect (Paquot, 2015). Besides providing methods and tools to assess the relative importance of words and their different uses, CL has introduced innovative approaches to lexicographical treatment of meaning, grammar and pragmatics, extending the scope of lexicographic research. Relying on Firth's contextual theory of meaning, evidence on the use of words in corpora has demonstrated to an unprecedented extent that words are not isolated, but combinations of syntagmatic patterns to acquire meaning.

Several questions that corpora can help answer cover the areas of language teaching and phraseology, the study of lexical bundles, collocations and frequently occurring preferred language sequences that are the central elements of corpus linguistics (Bennett, 2010, p. 8). 'Corpora are useful aids for phraseological enquiries as the language which falls between lexis and grammar is often not easily retrievable from grammars or dictionaries (Flowerdew, 2009, p. 408)'. To embrace lexis and grammar, an enhanced theoretical framework, construction grammar has emerged in recent years for phraseological studies in general, and especially for collocational analysis. The constructionist approach regards language as an idiomatic continuum in which constructions or building blocks can be defined as usage-based pairings of forms and semantic or discourse functions. Since collocations possess a construal nature with internal lexical restrictions, they are semantically predictable and can be described as symbolic units that span different phrasal patterns (Pastor, 2017, pp. 29-30).

3.6.2. Dictionaries as learning aids

The use of corpus data has spread rapidly in English pedagogical lexicography; in native-speaker, specialized and bilingual dictionaries the phraseology of words and lexical units is evidenced in corpora. Besides word lists, dictionaries are the basic reference sources for

vocabulary extension, either in printed or electronic forms, targeting learners with supposedly similar needs. Nevertheless, the focus has shifted from printed dictionaries to electronic dictionaries, according to Granger (2012, p. 2) 'Today lexicography is largely synonymous with electronic lexicography and many specialists predict the disappearance of paper dictionaries in the near future'. With the technological advances dictionaries tend to adapt to learners' lexicographic needs, one way of customization is to replace the static data of electronic dictionaries by articles containing dynamic data (Granger & Paquot, 2010, p. 3). Additionally, electronic dictionaries can be integrated directly into learning platforms; as an instance, with Hypertext function (http://www.lextutor.ca/hypertext) the user can click on a word in the submitted text to display the dictionary entry for that item (Boulton & De Cock, 2017).

The information provided in dictionaries on the authentic usage of a word in context or its frequent collocations is principally due to the development of corpus linguistics. Corpus-based dictionaries generally select vocabulary from a set of lexical patterns; the selection is based mainly on frequency information when defining an entry. A typical dictionary provides various kinds of information to the users, including meaning, pronunciation, part-of-speech, etymology and status (e. g. colloquial or standard). The authenticity of examples and the collocational patterns of an entry enhance disambiguation, illustrating even subtle changes in meaning and usage, and highlighting the grammatical properties of words. Nevertheless, dictionaries contain example sentences to illustrate the various meanings of words in meaningful context.

Monolingual dictionaries belong to three broad groups; *native-speaker dictionaries*, *learners' dictionaries*, and *technical or domain-specific dictionaries*. Non-native students tend to rely on advanced learners' dictionaries that focus on high-frequency words to greater extent than on the multiple meanings of words; this way, native speaker dictionaries and technical dictionaries can bridge the gap in text coverage. However, technical dictionaries also have their shortcomings; as they focus on a specific domain, they lack the vocabulary of academic English and the lexis of the adjoining subject fields (Kosem & Krishnamurthy, 2007, p. 2).

Thus far, three general collocation dictionaries are available in English, namely, the *BBI Combinatory Dictionary of English* (2009), *Oxford Collocation Dictionary for Students of English* (2009), and *MacMillan Collocations Dictionary* (2010). On the practical level of dictionary compilation Handl (2008) suggests that pairs of collocations should be profiled along lexical, semantic and statistical dimensions, and collocations to be included in a learner dictionary should be in the middle of this three-dimensional continuum. She developed a system

for identifying and selecting collocations, *Collocational Factor* (CF) statistics, applying CF to both the collocational direction of stronger and weaker constituents, as well as to the calculation of their weight. Handl (2008) proposed that a collocational entry of a learner dictionary should be presented under the entry of the stronger partner with a cross-reference of the entry of the weaker partner.

Beyond the high-frequency vocabulary, it is necessary to fulfil the specific needs and interests of learners when selecting words for study. Tourism dictionaries, mono- or bilingual, structured corpus-based databases, which contain touristic expressions and concepts besides tourism words, generally provide information not only on word meanings but on their etymology, and lexical and grammatical properties. Concerning electronic dictionaries their accessibility and speed are important; their greatest merit lies in their up-to-date information and continuous improvement. The only electronic *TourisTerm* dictionary is available in four languages at http://www.unwto.org/WebTerm6/UI/index.xsl, however, there is no information on the size of the dictionary. The website of *Word Tourism Organization* http://www2.unwto.org/ offers a *Tourism Term Glossary*, the entries provide brief yet highly professional definitions of the tourism terms (Terestyéni, 2009, pp. 397- 399).

3.6.3. Wordlists in vocabulary learning

Wordlists have proved to be useful guides in vocabulary acquisition, in the development of teaching materials, and in providing basis for language testing. There have been several attempts to compile word lists which facilitate foreign language learning, one of the oldest and most widely used is Michael West's (1953) *General Service List* (GSL) of over 2000-word families. Based on pre-computer corpus research the GSL was designed, combining objective and subjective selection criteria, to provide 'general service' vocabulary to foreign language learners. However, over a couple of decades it was criticized for being outdated and based on a small corpus compared to the standards of electronic corpora. For this reason, on the sixtieth anniversary of the GSL, the modern computer-based *New General Service List* (Browne, 2013, p.13) was introduced that was created on the basis of the 1.6 billion-word *Cambridge International Corpus (CEC) of spoken and written texts*. The *University Word List* (UWL) (Xue and Nation, 1984) consists of 836 general academic words which are not included in the GSL, but 'frequent and of wide range in academic texts' (Nation and Waring, 1997, p. 17). The *Academic World List* (AWL) (Coxhead, 2000), the *Academic Formulas List* (Simpson-Vlach

& Ellis, 2010), the *new Academic Vocabulary List* (AVL) (Davies, 2012), and instruments, like the *Vocabulary Size Test* (Nation & Beglar, 2007), were compiled using word family as a grouping principle. Yet, these word lists of academic English have one major flaw, they exclude high frequency words, especially the first 2000 words of the *General Service List* (West, 1953), assuming that students should already know these words (Kosem & Krishnamurthy, 2007, p. 2).

Recent publications (Brezina & Gablasova, 2017; McLean 2017; Stoeckel, Ishii, & Bennett, 2018) concern with the principles of word list development; that is, which words should be included in such lists based on selection criteria. In the aforementioned word lists the headwords are grouped with their inflections, regular derivations, and frequent irregular derived forms. An important question is, if the learners are aware of the meaning of the headword, how many related words are they supposed to know. However, there are arguments that word families are not the appropriate forms for language teaching and assessment (Schmitt & Meara, 1997; Gardner 2007). Grounded on the findings of a recent study McLean (2017) concluded that most advanced learners had limited knowledge of derived forms, thus, he proposed the introduction of *flemma*, a less inclusive modified lemma, as the basis for word lists. A lemma consists of a headword and the same part of speech inflections, whereas in a flemma all parts of speech inflections are included. Yet, flemma often proved to be too broad; Stoeckel *et al.* (2018) found that in cases where learners demonstrated their lexical knowledge in one part of speech, they understood the word in another part of speech only 56% of the time.

3.6.4. Vocabulary selection and text coverage

In direct vocabulary teaching the selection of words to be studied is one of the main concerns of foreign language teachers. A key feature of the lexicon is the fact that a relatively small proportion of words in a language are highly frequent, accordingly, decisions about vocabulary teaching and learning are generally made on the basis of frequency. Obviously, there is a payoff of concentrating initially on the most frequent words since they account for a high percentage of running words in written and spoken texts. As there is no objective threshold, the question of how much vocabulary a second language learner needs remains to be answered on the basis of frequency (Read, 2004/b, p. 148).

Goulden, Nation and Read (1990) observed that the most significant difference between native speakers and second language learners seems to be in the size of their lexicon. To select and

control the lexical items that are most relevant to teach, language teachers need to encompass the required core vocabulary that enables an individual to communicate at a certain proficiency level. Besides the general core vocabulary, the terminological lexical minimum of the essential specialised vocabulary also needs to be outlined.

As discussed in 2.6.1. Chang and Nation (2004) differentiated four levels of lexis acquired in language courses; (1) high frequency or *general lexis*, (2) *academic or semi-technical lexis*, (3) *technical*, and (4) *low-frequency lexis*. These categories of lexis can be delineated in terms of frequency, range, keyness or text coverage. Text coverage can be defined as instances of words, or percentage of tokens that coincide with the constituents of a particular word list.

Schmitt (2000, p. 142) estimated that learners of English need to know at least 2000 frequently used words for successful everyday conversation and to understand the gist of a text. A vocabulary size of 2000 words provides nearly 80 per cent coverage of written texts, that is, one word in every five is unknown (Nation & Waring, 1997, p. 9). Research by Liu Na and Nation (1985, p. 37) showed that this ratio of unknown words does not allow successful inference of meaning, at least 95 percent coverage, approximately a vocabulary of 3000 words, is needed for the reasonable comprehension of an unsimplified text. Following a similar approach, Schmitt and Schmitt (2014) proposed a 3,000 high-frequency word threshold to provide satisfactory text coverage, yet, as previously Laufer (1997, p. 144) had claimed, the 95% text coverage can be achieved only with a 5,000-word vocabulary or with knowing 3,000 word families. Nevertheless, Hu and Nation (2002, p. 422) found that some learners need to know about 98 % of the words to read independently, thus the vocabulary threshold for listening and reading comprehension needs to be set higher than the widely recommended 95 % coverage. In spoken language, however, about 1,800 words make up over 80 per cent of a spoken corpus (O'Keefe et. al, 2007, p. 32), and a 2000-word vocabulary is sufficient to gain approximately 95% coverage in conversational discourse (Adolphs & Schmitt, 2003). As the 95% figure represents a considerable number of unknown words, Adolphs and Schmitt (2003) suggest that a 3000-word vocabulary is needed for learners to successfully perform everyday conversations.

Students who pursue their studies in English need to have an extensive vocabulary to develop their academic writing, literacy and communicative skills. According to Nation and Waring (1997, p. 7) native university graduates have a vocabulary size of up to 20,000 word families by adding roughly 1000 word families a year to their lexicon. Nation (2006, p. 78) maintained

that if 98% is taken as the ideal coverage, for a university student at least an 8,000-9,000 word-family vocabulary is needed for dealing with written text such as academic or specialized texts, and 6,000-7,000 families for dealing with spoken texts. This size of vocabulary constitutes 570 word families listed in Coxhead's (2002) *Academic Word List* (AWL), in addition to the 3,000 high-frequency words. Although a large number of second language learners can achieve native-like vocabulary size, many adult foreign learners have a vocabulary size less than 5,000-word families after having studied English for several years.

Beyond the high frequency first 2000- and 3000-word families, it is necessary to take into account the vocabulary needs of specific groups of learners when selecting lexical items to study. To cater for the demand in teaching English for academic and occupational purposes worldwide, particular attention has been paid to specific vocabulary; one way of vocabulary approach is to focus on frequently occurring academic words, the other is to identify sub technical terms related to particular disciplines. The creation of small corpora comprising discipline-related texts is a well-tried method in teaching English for specific purposes, whereby advanced students access in-house corpora to investigate the frequency and meanings of particular domain-specific lexical items (Read, 2004/b, p. 151). Gavioli (2002) acknowledges the didactic value of such corpus explorations, but also warns that learners should not make generalizations on the usage of words without confirming it with evidence from a larger general corpus.

3.7. Concluding remarks

In this chapter some ways are presented how corpus linguistics and corpora can promote language learning and teaching, and highlighted that corpus-based didactic approaches offer valuable resources in terms of ESP teaching materials, and corpus-based vocabulary teaching and learning methods. Concordances can be applied as complementary to teaching materials, in the design of syllabi and creating word lists for pedagogical purposes. While working with corpora students observe lexical and grammatical patterns, as well as nuances in their usage, gain information about language features, furthermore, this 'search-and-discovery' process not only facilitates language learning but also promotes learner autonomy.

Corpus linguistics as a research approach has supported empirical studies of language variation and use, and with the greater feasibility and generalizability of corpus-based investigations it

has reoriented our approach to language teaching in the past few decades. One of the greatest advantages of corpus linguistic methodology is that it provides language researchers with a large amount of naturally occurring language samples, allowing thorough analysis of the nature and behaviour of lexical patterns in different registers. Corpus-based research has also highlighted that lexis and grammar are interdependent and interact to create meaning, thereby it has fundamentally changed our notions of what it means to know a word. Owing to corpora we have developed new insights into the structure of language and gained better understanding of contextual language use, and, as a result of this, corpus linguistics has reshaped our views of language acquisition.

THE TOURISM ENGLISH CORPUS AND THEMATIC WORD LISTS

4.1. Introduction

This chapter explicates the rationale of word lists for vocabulary teaching and learning, gives an overview of the main aims, aspects, stages and procedures of the development of the Tourism English Corpus (TEC), and presents the methods and compilation process of thematic wordlists on the main fields of tourism. After having set forth the results of a pilot study on vocabulary profiling and wordlist compilation, the findings of keyword and frequency analyses are discussed, delineating the most important lexis of the main domains of tourism on the basis of six topic-related subcorpora. The findings of a brief analysis on text coverage, range and distribution of some key terms are presented, and the pedagogical implications of thematic word lists in tourism classes are concerned.

4.2. Word lists in vocabulary teaching and learning

Since Meara's statement in 1980 pertaining to the neglected status of vocabulary teaching and learning the situation has changed considerably, significant pedagogical research related to vocabulary acquisition has been carried out in the past decades (discussed in 2.1.). However, as Folse (2010, p. 139) observed, 'explicit vocabulary focus is surprisingly low' in English language classes, despite the fact that learners need to master vast amounts of general and subject-specific words. General and specific word lists as key resources offer great potential in vocabulary teaching and learning; with explicit teaching of words and related vocabulary learning strategies students can extend their lexicon, and use vocabulary items more effectively. The most well-known English vocabulary lists, the *General Service List*, the *New General Service List*, the *University Word List*, and Coxhead's (2000) *Academic Word List* cover the most frequent common English words and academic vocabulary (word lists and text coverage

are discussed in 3.2.4.1. and 3.5.4.), these word lists can help learners to fill the gaps in their vocabulary knowledge.

Although some teachers prefer teaching words in sentence context and consider explicit vocabulary learning from word lists and glossaries a kind of rote learning, studies revealed (Laufer & Shmueli, 1997; Griffin & Harley, 1998) that words presented in lists promoted retention to a greater extent than the explanation of words in English, or their oral presentation in context. As Folse (2010) claims, 'there is practically no evidence to suggest, that learning new words from lists is in itself detrimental '(p.40), therefore he advocates the use of vocabulary lists, yet warns that teachers should not rely exclusively on them without taking into account students' learning preferences.

Word lists can be the starting points of course design, they help language teachers decide what vocabulary to include in the syllabus/lexical syllabus. Nation (2016) highlighted the benefits of using vocabulary lists in language teaching;

> Word lists lie at the heart of good vocabulary course design, the development of graded materials for extensive listening and reading, research on vocabulary load, and vocabulary test development (xi).

Nation (2001, 2007) maintains that a well-planned course has four equal strands of *meaning-focused input*, *meaning-focused output*, *language-focused learning*, and *fluency development*. Language-focused learning includes explicit vocabulary teaching and learning, thus frequency-based word lists can be used as checklists for vocabulary extension, taking into account learners' receptive and productive word knowledge.

Attentive course design focuses on high-frequency words, proceeds to mid-frequency words, and pays attention to non-repeated words that support language learning. In addition to *frequency*, *range* and *distribution* are measures of the usefulness of a word. *Range* refers to the number of occurrences in different texts; a word that is frequent in one text but does not occur in other texts has a very narrow range. *Distribution* measures how evenly the words occur in different texts, thus, a word that occurs with equal frequency throughout the texts has a high dispersion score. Word lists based on frequency, range and distribution are useful guides of systematic vocabulary selection, and for preparing listening and reading materials.

Besides vocabulary development, word lists are useful aids in language testing; the construction of vocabulary size tests, and the creation of tests assessing particular levels of vocabulary can

rely on well-balanced lists. Vocabulary tests based on corpus-derived word lists are useful in setting course objectives and in monitoring learners' progress. The word lists used in testing should reflect the learners' goals for language use, as there is no sense in using vocabulary lists generated from speech corpora in a course that develops writing skills (Nation, 2016).

4.3. Tourism English word lists

In English for tourism classes it is indispensable to provide appropriate vocabulary related to a range of tourism sectors. Although there are several tourism-related textbook glossaries and dictionaries, they narrowly cover the specific needs of different tourism occupations. There have been some enterprises to create tourism word lists, but unfortunately none of them is available publicly. Scott and Revell (2004) compiled a wordlist for the hotel and catering industry, whereas Chujo, Utiyama and Oghigian (2006) extracted a tourism vocabulary wordlist from a large-scale corpus that contains texts related to inbound tourism. With various statistical measures the words were categorized into three levels of difficulty, however, the list focused only on one subfield of the travel industry.

Fujita (2009) attempted to compile a more comprehensive English in tourism wordlist for university tourism students that encompassed four main domains, (air travel, tour operators, accommodation, and tourism-related academic journals), based on a self-built corpus comprising 40,000 tokens of written texts. The four subfields were selected on the grounds of earlier needs analysis in a tourism programme. A keyword list was generated for each subfield, and, besides the four keyword lists that consisted of rather different sets of words, a combined list of the most characteristic 150 words was created.

Fujita (2010) further investigated the extent of overlap between tourism industry vocabulary and tourism academic vocabulary based on an extended corpus of written texts that comprises 90,000 running words. It was found that there was only 14% overlap in these two subfields, indicating that the vocabularies of these tourism fields are highly distinct from each other. The results suggested that teachers should select from and combine these lists, adapting them to students' specific needs.

4.4. The Tourism English Corpus

4.4.1. The rationale of the compilation of the Tourism English Corpus

The *Tourism English Corpus* (TEC) was created with the aim that it would provide basis for the compilation of pedagogically applicable wordlists and collocation lists, and for the investigations of tourism lexis and discourse. Since three subcorpora had been investigated in previous studies, it was corroborated that as few as approximately 13,000 words are adequate for compiling a pedagogically applicable word list, 30,000 running words allow investigations into the lexico-grammatical patterns of tourism language and peculiarities of tourism discourse, and 50,000 tokens are sufficient for collocational analysis and for the compilation of a collocation list. In corpus design the constraints of time, the availability of texts, and copyright issues demanded to make compromises concerning the size and structure of the corpus. The corpus has its limitations, yet, it proved to be adequate for the examination of the pre-planned issues, namely, for the compilation of word- and collocation lists, and for the inquiry into the specific lexical and syntactic patterns of tourism language and discourse.

The corpus-based investigations focused on the following questions:

1. How do the thematic wordlists reflect the lexis of that particular tourism subfield?
2. What is the frequency and distribution of subject-specific words in the TEC?
3. What coverage do the TEC wordlists provide for reading comprehension?
4. What are the potential pedagogical implications of the findings?

4.4.2. General design of the corpus

Applying Jablonkai's (2010, p. 118) *Model for corpus creation for ESP* and her proposed steps of corpus compilation (as discussed in 3.6.4.), both theoretical issues such as structure, size and sampling, and practical aspects including data collection, storing and copyright issues were considered in the design of the *TEC*. Pertaining to the practical aspects of corpus design, copyright issues designated the means of data collection; only extracts of travel articles, product information and customer reviews that comprise fewer than 400 words were included, except for the subcorpora investigated in the preliminary studies whose texts were downloaded from a free website. Since the main reason of the unavailability of tourism corpora lies in the fact that it is impossible to handle the copyright issues of the numerous websites, the *TEC* comprises

only unidentifiable excerpts of tourism texts. Nonetheless, due to legal constraints, the corpus can be used exclusively for research purposes (cf. Meyer, 2004, p. 78).

Having defined the aim of the corpus, that is, providing basis for lexical, syntactic and discoursal analyses of tourism language, a sampling frame was established to achieve optimum representativeness. To define the structure of the corpus several aspects had to be considered; selecting the relevant text categories and defining their proportion to create an optimally balanced corpus, as well as covering the diversity of the topics. The size of the corpus was pre-established by the previously built subcorpora; all but two are made up of approximately 30,000 tokens, as this number of running words had proved to be sufficient for lexico-grammatical and discoursal analyses in preliminary studies. Besides legal issues, availability and feasibility had to be taken into account in the phase of data collection, and, as a final step, the texts were converted to plain text files and saved in electronic format.

To select and evaluate the relevance of the texts to be included, a quick assessment of their content was carried out. Pearson (1998, p. 54) suggested that relevancy could be determined 'by looking at what a particular text is about (e.g., based on a title, or headings), and 'by examining the lexical structure of a text and identifying keywords used frequently in the text'. Following this approach, the texts were selected on the grounds of tourism terms that I considered keywords in that field, and having assessed the content, only the most relevant parts were included in the corpus.

4.4.3. The structure of the Tourism English Corpus

Six subcorpora were combined to constitute a tourism corpus, they were created with the same design principles, they have the same structure, and each subcorpus covers a main sector of tourism, including air travel, different types of holidays and services. To ensure representativeness in the main fields of tourism the texts were selected from frequent and accessible text categories. The subcorpora contain passages of travel articles, product information and customer reviews that differ in length, the number of tokens ranges from 13,582 to 50,029 in the subcorpora. *Table 1* presents the structure of the *Tourism English Corpus*.

Table 1: *The subcorpora and the number of tokens in the Tourism English Corpus*

THE TOURISM ENGLISH CORPUS	
Subcorpora	**Number of tokens**
air travel	13,582
beach holidays	50,029
city tours	30,846
hotels and accommodation	30,105
restaurants and catering	30, 077
wellness and spa tourism	30, 108
Total:	184,747

The difference between the number of running words is due to the fact that three subcorpora, the *air travel*, the *beach holidays* and the *city tours* were compiled previously to examine various features and lexical properties of tourism language. Although the different sizes of subcorpora might give a hint of imbalance, the strictly subject-related texts are representative of the core vocabulary of that domain. The *Tourism English Corpus* contains a total of 184,747 running words, and can be later extended with subcorpora based on needs analyses with students' cooperation. The TEC corpus is available online at http://bit.ly/43ZHyfA.

Preceding the compilation of thematic word lists generated from each subcorpus, a pilot study was carried out (Kiss, 2012) to investigate whether a small-size self-made specialised corpus, comprising as few as approximately 13, 000 tokens, is adequate for creating a proper thematic word list.

4.5. A pilot study: The vocabprofile and keyword list of the air travel subcorpus

4.5.1. Reasons for the study

Teachers of ESP often experience the difficulties of vocabulary teaching and assessment; besides staying objective it is difficult to make a decision on the relevance of vocabulary items to be taught and tested. To investigate whether the most important lexis of a topic can be encompassed with the help of vocabulary profiles and frequency lists, I compiled a small-scale corpus and examined the potentials of utilizing the findings in vocabulary teaching and assessment. My assumption was that with a representative amount of data I would be able to

select the most important and most frequent words of special lexis in specific texts, and aimed to find evidence that even a small-scale corpus provides sufficient amount of data to find recurrences of genre-specific words in tourism context.

4.5.2. Compilation of the air travel subcorpus

To develop my corpus, which was incorporated as a subcorpus in the *TEC*, the articles on *air travel* were downloaded from on-line tourism magazines and websites. As tourism is a rather broad issue, I had to encompass a narrower scope and focused on this major topic. Since corpus design requires careful selection of texts on a particular topic, bearing in mind what the corpus is designed for, I collected extracts from twenty-six professional travel articles that comprise approximately thirteen thousand tokens, and saved them in plain text format. Afterwards a quantitative frequency analysis was conducted assisted by the processing tools of the *Compleat Lexical Tutor* (CLT), in order to find recurrences to include in a pedagogically applicable vocabulary list. Specialized corpora have considerable advantages; as the texts are systematically selected, they are likely to represent the target domain more reliably than randomly chosen texts, and the specialised lexis is more probable to occur even with a smaller amount of data. Furthermore, it is likely to be easier to define how they can be applied didactically. The subcorpus dubbed *air travel* is available online at http://bit.ly/GD4fHL.

4.5.3. Corpus tools

The tools used for analysing the *Vocabprofile* (VP) of the corpus can be found on the *CLT* Web site, *www.lextutor.ca.*, which is divided into three sections; *Tutorial*, *Research*, and *Teachers*. A powerful tool of the Research section, *Vocabprofiler*, breaks texts down by word frequencies and divides them into one, two, three or five thousand levels, academic words, and the remainder, or 'offlist'. The other research tool which proved to be useful for lexical analysis is *Keywords*. The keywords are those words that are far more frequent proportionally in the articles than in a general reference corpus; the reference corpus is the *Brown Corpus* whose one million words comprise 500 texts of 2000 words on a broad range of topics.

4.5.4. Findings of the study

The *Vocabprofiler* not only demonstrates the frequency of words at different levels, but classifies the items as types, and a numerical count is made on the number of tokens. The type-token ratio demonstrates the lexical diversity of the text as can be seen in *Table 2*. Low TTR

(type-token ratio) indicates that the text is lexically poor with high recurrence of certain words, the more types a text comprises the more sophisticated style it has. I assumed that in specialised texts the type-token ratio is lower and should be compared to those of in other texts at different levels. With the help of the *Web Frequency Indexer,* I was expecting to find the most significant vocabulary items of professional texts. However, it did not reveal the results hoped for as the proportion of function words in journalistic texts is high, and there were only two content words, *airlines* and *flight,* among the first twenty most frequent items.

Table 2*: CLT's Vocabulary Profile output (www.lextutor.ca/vp)*

Words in text (tokens):	13 582
Different words (types):	2620
Type-token ratio:	0.19
Tokens per type:	5.18
Lexical density (content words /total)	0.58

Examining the results of VP my assumption on the low type-token ratio of specialised texts has been verified as the number of tokens per type is relatively high. However, the lexical density, the number of content words compared to the number of tokens, seems higher than it was expected after having tried the *Web Frequency Indexer*.

Table 3*: CLT's Vocabulary Profile output (www.lextutor.ca/vp)*

	FAMILIES	TYPES	TOKENS	PERCENT
K1 words (1-1000):	669	1155	*9836*	*72.42%*
Function:			5717	42.09%
Content:			4119	30.33%
K2 words (1001-2000):	257	357	*1158*	*8.53%*
1k+2k				(80.95%)
Academic words:	232	344	*725*	*5.34%*
Off-list words:		767	*1863*	*13.72%*

Table 3 provides details on the VP that demonstrates the occurrence of words at different levels. It reveals that slightly more than 80 percent of the tokens belong to K1 and K2 levels, and despite the professionalism of the selected articles the rate of academic words is relatively low. The significant number of off-list words refers to the specialized lexis and the journalistic style of the texts, and it might indicate the high number of proper nouns and hapax legomena, words that occur only once in the corpus.

Table 4: *CLT's Keyword output (www.lextutor.ca/keywords)*

(1) 2492.00 airline
(2) 1691.40 airlines
(3) 1510.00 carry-*on*
(4) 1057.00 check-*in*
(5)   944.00 baggage
(6)   755.00 fliers
(7)   680.00 online
(8)   604.00 round-*trip*
(9)   566.30 luggage
(10) 529.00 airbus
(11) 453.00 cancellations
(12) 453.00 skybus
(13) 378.00 rebooking
(14) 345.00 delta
(15) 339.75 airports

As *Table 4* demonstrates the *Keyword* list corroborates that the systematically selected texts of a relatively small corpus provide reliable results, and the most frequent content words are characteristic to the specialised lexis of aviation. The number accompanying each word indicates the number of times more frequent the word is in the texts than in the *Brown Corpus*. The first item of the output, *airline*, has 3 natural occurrences in the *Brown Corpus* and 99 occurrences in the *air travel* subcorpus.

4.5.5. Conclusion of the pilot study

The results suggest that corpus-derived evidence on the use of special lexis can be applied to the acquisition and assessment of specialist vocabulary, and even a relatively small-sized corpus provides sufficient amount of data for generating proper wordlists. The analysis of recurrences

can be extended from single words to lexical bundles and collocations, and the results and their implication can be utilized in vocabulary-based tasks. Furthermore, the texts of the corpus can be either compared to students' textbooks, or used in reading comprehension or gap-filling activities.

4.6. Creation of tourism word lists based on the TEC

4.6.1. Unit of text analysis

The compilation of tourism wordlists was guided by Thomas Cobb's clear statement; 'Learners like word lists, so let's give them good ones, plus something useful to do with them' (https://www.lextutor.ca/list_learn/). First, frequency lists of subcorpora were created and then organized into word families as units of analysis. The next step was the generation of keywords lists from each subcorpus; finally, words were selected to include in the vocabulary lists from both frequency and keyword lists, involving intuition as well. For validating and fine tuning the word lists *Range* programme was used, as it can run up to 10 base word lists, this way handles the lists of *TEC* and the subcorpora, respectively.

A word family is constituted by *flemmas*, that is, headwords and their all parts of speech inflections (see details of units of analysis in 4.3.2. and 4.5.3.). The assumption that understanding a derived form of a word family does not require extra effort from language learners, if they are familiar with base forms and affixes, was confirmed by empirical evidence. Schmitt and Zimmerman (2012) found that non-native English-speaking students showed increased knowledge in noun and verb derivatives, however, adjective and adverb forms seemed to be somewhat more difficult for them. Coxhead's (2002) research revealed that word family is a psychological unit of the mental lexicon that helps create interconnected links of words, thus the reinforcement of word stems promotes retention. Some instances of word families of the *BNC* are presented in *Table 5*.

Table 5*: Instances of word families in the British National Corpus (BNC)*

Headword	Members of the word family
CLASS	*classed, classes, classing, classless*
DETAIL	*detailed, detailing, details*
EFFECT	*effected, effecting, effective, effectively*

However, there are some limitations of the application of word families; it can be difficult to decide which word forms should be considered members of a word family, and it can be problematic to classify the remaining words into word families.

4.6.2. Corpus processing tools

Three freely available corpus processing tools were used to create and analyse wordlists depending on the practicability of their functionalities: *Antconc 3.5.8*, the initiative *Compleat Lexical Tutor*, and *Range* programme. The utility of the *CLT* tools to investigate the VP of the air travel subcorpus had been attested in the pilot study. *AntConc 3.5.8* developed by Laurence Antony is a relatively simple freeware concordance program including several powerful tools that can be run on or downloaded from the website *http://www.antlab.sci.waseda.ac.jp/antconc_index.htm*. It can be used effectively for processing the data of a corpus comprising relatively high number of tokens, as it saves the uploaded text files. Its toolkit includes the *Concordance, Concordance plot, File view*, *Clusters, Collocates, Word List* and *Keyword List* functions to generate concordance lines, analyze word frequencies and lexical bundles.

Range programme that was designed for Windows based PCs is available at Paul Nation's website http://www.vuw.ac.nz/lals/staff/Paul_Nation. Range can be used to compare the lexis of texts in up to 10 files. It can also be used to generate word lists based on frequency and range, and to display the coverage of tests compared to available base lists. These lists include the most frequent 1000 words in English, the second 1000 most frequent words, and a third base list comprising words that are not in the first 2000 words, but frequent in upper secondary school and university texts covering a range of subjects. The lists are based on the *GSL* (West, 1953), the *AWL* (Coxhead, 1998, 2000), and contain other commonly used words such as months, days of the week, and numbers, and include both English and American spelling.

4.6.3. Steps and procedures of thematic word list compilation

First, thematic word lists were created on the basis of each subcorpus, and then a concise TEC word list was compiled from these lists, the compilation of the TEC list is presented in 5.6.6. The individual steps and procedures of establishing the *thematic* wordlists are as follows:

1. Generating frequency lists of the TEC by *CLT Frequency List Builder* function
2. Organising the content words of frequency lists into word families applying the lemmatizing function of *Compleat Lexical Tutor*
3. Creating keyword lists with the help of *Antconc 3.5.8*
4. Selecting and storing words from both frequency and keyword lists
5. Validating the final lists with *Range* programme.

4.6.4. Frequency of words

4.6.4.1. Frequency of words in English texts

Words in a language do not occur with similar frequency; some words like *the, a/an* or *of* occur in almost every sentence throughout a text, while students might meet some words only once or not at all during their studies. The occurrence of words in a 1000-word text or above, or in a collection of texts, can be delineated by Zipf's (1935, quoted in Nation, 2016, p. 9) law that asserts that *rank* multiplied by *frequency* yields a constant figure as we go downwards a frequency-ranked word list. However, the frequency of words in such lists decreases very quickly, about halfway down the words tend to occur only once in the text.

A relatively small group of words have fairly high occurrence, the ten most frequent words cover around 25% of running words in written English texts. The most frequent 100 English words cover nearly 50% of the running words in a text, and the first 1000 different words provide text coverage between 70% and 90% depending on the content, or whether they occur in written or spoken texts. However, a large group of words have low frequency of occurrence, and the boundary between high frequency and low frequency words is arbitrary. Traditionally, the first 2000 most frequent English words are called high-frequency words, up to 6000 mid-frequency words, and from 7000 low-frequency words. Evidently, high-frequency words need to be in the focus in course design, considering the cost/benefit principle of vocabulary learning (Nation, 2016, pp. 3-5).

4.6.4.2. Frequency lists of the TEC

As a first step, frequency lists of the TEC were created with the help of *CLT Frequency List Builder,* and then confirmed with the *Frequency* application of *Range* programme. The *CLT Frequency List Builder* indicates the rank in a frequency ordered list, the raw frequency of words, and their individual and cumulative text coverage. The frequency lists of *Range*

programme, besides rank and the number of occurrences provides information about the ratio of occurrence. *Table 6* shows the output of the first 10 most frequent words of the TEC in the *CLT* list, the first 100 most frequent words listed by *CLT* are given in APPENDIX A.

Table 6: *Output of the 10 most frequent words of TEC generated by CLT Frequency List Builder*

RANK	FREQUENCY	COVERAGE		WORD
		individual	*cumulative*	
1.	11172	6.06%	6.06%	**THE**
2.	7173	3.89%	9.95%	**AND**
3.	4941	2.68%	12.63%	**A**
4.	4599	2.49%	15.12%	**OF**
5.	4397	2.39%	17.51%	**TO**
6.	3001	1.63%	19.14%	**IS**
7.	2938	1.59%	20.73%	**IN**
8.	2173	1.18%	21.91%	**FOR**
9.	2038	1.11%	23.02%	**YOU**
10.	1757	0.95%	23.97%	**WITH**

In the sample, the word type <u>and</u> is the second most frequent word, it occurs 7173 times in the corpus, and with the most frequent item, <u>the</u>, it covers cumulatively 9.95% of the texts. On its own it covers 3.89% of the texts, that is, 9.95% minus 6.06%. As it is demonstrated, there was no content word among the first ten most frequent items, and the top 100 frequent word list contains only 33 content words, including 21 nouns (NUMBER, BEACH, CITY, SPA, PLACE, HOTEL, FOOD, EXPERIENCE, SERVICE, ROOM, TIME, STAFF, TOUR, MASSAGE, AREA, RESTAURANT, WATER, ISLAND, WORLD, DAY, PEOPLE). In addition to nouns, there are 7 adjectives (GREAT, BEST, BEAUTIFUL, FRIENDLY, PERFECT, NICE, AMAZING), 3 main verbs (GET, VISIT, ENJOY), and 2 adverbs (VERY, REALLY) in the list.

Validating the frequency lists, the *Range* programme provided more details about the occurrence of words. *Range* not only counts the frequency of word types in different files, but also records the total frequency of occurrence of individual word types in each file, as well as counts the number of files in which a particular word type occurs. *Function text* base word list of *Range* software helps eliminate frequently occurring function words from the generated

vocabulary lists, this way only content words are selected. Besides omitting function words, it became obvious that a frequency threshold needed to be set to decide what important vocabulary items should be included in the finalised word lists. *Table 7* shows the frequency output of a few examples from the TEC, sorted by *Range* programme. Evidently, these target words have the highest occurrence in the subcorpus they are characteristic to; accordingly, 'food' is the most frequent in the *restaurants and catering* subcorpus.

Table 7: Sample output of frequency (FREQ) in Range programme

WORD TYPE	TOTAL FREQ	FREQ air travel	FREQ beach holidays	FREQ city tours	FREQ hotels & accommodation	FREQ restaurants & catering	FREQ Wellness & spa
beach	725	2	583	14	113	8	5
city	618	5	65	461	40	40	7
place	504	6	129	68	67	116	118
hotel	479	8	31	37	366	5	32
food	465	8	33	27	13	367	17

After generating the frequency lists, the second main step was the grouping of content headwords with their word families comprising all the inflectional variants/ flemmas. The lemmatizing function of CLT proved to be the most feasible solution for data processing. Samples of headwords of the *hotel and accommodation* subcorpus can be seen in *Table 8.*

Table 8: Samples of headwords and word families from the hotels and accommodation subcorpus

HEADWORDS	WORD FAMILIES OF RELATED WORD FORMS
accommodate	*accommodated, accommodating, accommodates*
buffet	*buffets, buffeted*
discount	*discounted, discounting, discounts*
disturb	*disturbs, disturbed, disturbing*
floor	*floors, floored, flooring*
housekeeper	*housekeepers, housekeeping*
lodge	*lodged, lodges, lodging*
overlook	*overlooked, overlooks, overlooking*

To illustrate, in the output of *hotels and accommodation* texts generated by *CLT Familizer/lemmatizer* 2,934 headwords were extracted and identified from among 30,455

running words. There were 235 (8.01%) unclassified word types including *amenities* or *airbnb*, however, the majority of these words are of foreign origin such as *volkshotel* or *hüttenpalast*.

Four selection criteria were used in the compilation of wordlists based on frequency. Firstly, only content words with high occurrence that are characteristic in that particular subfield were included in the lists. Secondly, a frequency threshold was set for the inclusion in the lists; types with more than 10 occurrences were selected for the full TEC word list, and with higher than 5 occurrences in the thematic wordlists elicited from the subcorpora. Thirdly, low frequency but highly specific words were sorted out to include, based on teacher intuition. Range was considered less significant than frequency in the selection process, as the thematic word lists represent the vocabulary of a particular tourism subfield. The final selection criterion was the occurrence of specific word families, however, that member of the word family was included in the word lists that seemed to be most likely to occur in that context.

As demonstrated in *Table 6* and had been confirmed in the pilot study, due to the prevalence of function words a frequency list in itself does not contain and represent the most distinctive content words of that particular domain, therefore, frequency and keyword lists need to be merged to generate applicable thematic word lists.

4.6.5. Generating keyword lists

For the selection of keyword items Pojanapunya and Todd's (2016, pp. 3-10) current approaches were applied; following the selection of top N items a high frequency item threshold was set. However, there are drawbacks of setting high frequency thresholds; items with very large frequency differences, as well as important low frequency words may be excluded. To eliminate these drawbacks all items were manually examined and regarded as candidate key items in the finalised vocabulary lists. Therefore, according to Gabrielatos (2017), 'keyness is not a straightforward attribute, and a keyness analysis does not necessarily entail objectivity'.

According to Chung and Nation (2004) technical lexis can be identified best by comparing frequency data in a general and a specialised corpus; words that occur more often in specialized texts than in a general corpus can be regarded as technical vocabulary. For the generation of keyword lists the keyword function of *Antconc 3.5.8* was used, the running words of the TEC were compared to a general reference corpus, to the BNC combined list of written and spoken

texts. The keyword list function sorts and ranks the keywords based on frequency and keyness, the statistical measure applied for generating keywords is the Log-likelihood ratio, and the keyword statistic threshold is set as p< 0.05. The total number of tokens of the TEC is 184,747, the total number of types is 13,886, and 1,536 words were identified as keywords. *Table 9* shows the first twenty content keywords of the TEC and its subcorpora, respectively.

Table 9: *The first twenty content keywords of the TEC and its subcorpora generated by Antconc 3.5.8*

			THE FIRST TWENTY CONTENT KEYWORDS			
RANK	TEC	AIR TRAVEL	BEACH HOLIDAY	CITY TOURS	HOTELS ACCOMM.	RESTAURANT CATERING	WELLNESS SPA
1.	*place*	airline	*beach*	tour	*hotel*	*food*	*massage*
2.	*great*	carry-on	island	*city*	*room*	*service*	*spa*
3.	*beach*	check-in	vacation	see	very	very	*experience*
4.	*city*	baggage	water	world	stay	restaurant	treatment
5.	*food*	fliers	enjoy	visit	*great*	*great*	relaxing
6.	*good*	on-line	world	*beautiful*	pool	*good*	*good*
7.	*experience*	round-trip	*place*	*best*	clean	*experience*	*place*
8.	*best*	luggage	perfect	history	*beach*	*staff*	*staff*
9.	*hotel*	airbus	*best*	*place*	*area*	menu	*time*
10.	*service*	cancellation	tourist	travel	nice	delicious	*area*
11.	*time*	skybus	dive	attraction	*staff*	excellent	*great*
12.	*room*	rebooking	popular	offers	breakfast	wine	*room*
13.	*spa*	airport	activities	*great*	location	atmosphere	pool
14.	*area*	passenger	*beautiful*	*time*	*friendly*	*friendly*	sauna
15.	well	plane	visit	enjoy	free	dining	*friendly*
16.	*massage*	fee	*area*	people	*place*	amazing	body
17.	*staff*	ticket	offer	popular	guests	*best*	relaxation
18.	like	air	holiday	sights	comfortable	dish	*best*
19.	*beautiful*	flying	sports	tourists	restaurants	nice	day
20.	*friendly*	connection	sun	famous	walk	meal	amazing

The keywords of the TEC that are also key items in the subcorpora are italicized; the number of their occurrence as keywords across the subcorpora is indicated in the parentheses. The most frequent key nouns in the subcorpora are PLACE (5), ROOM (3), AREA (3), STAFF (3), EXPERIENCE (2), TIME (2), and BEACH (2). Frequently occurring adjectives include FRIENDLY (3), GOOD (3), BEST (3), and BEAUTIFUL (2), however, there are key adjectives in the subcorpora that are not in the TEC list, such as AMAZING (2) or POPULAR (2). The

words PLACE, BEACH, CITY and FOOD were the highest frequency content key nouns in the TEC list, indicating their overrepresentation and importance in tourism texts.

It is interesting to take a look at the negative keywords, that is, to words that are significantly less frequent in the *TEC* than in the *BNC* as a reference corpus. Negative keywords are mainly items of general lexis; the majority of negative keywords are not tourism-related. These general words include a long list of proper nouns, adjectives referring to nations such as PORTUGUESE, JAPANESE and CUBAN, descriptive adjectives including SEASONAL, SPLENDID, and OUTSTANDING, connectors as HOW, WHILST or PLUS, and non-words in English *e.g.*, LE.

4.6.6. The compilation of the full TEC word list based on the thematic word lists

As it was attested, frequency exclusively does not signify the most specific lexis of tourism, keyness and keywords also need to be taken into consideration when generating word lists. Since *Antconc 3.5.8 Keyword list function* enumerates the keywords in the order of frequency, it seemed to be the most feasible and reasonable way of selecting some headwords, and choosing vocabulary items from *Antconc Wordlist* that displays the words in alphabetical order. The final selection criteria to the compilation of the full *TEC wordlist* were defined as follows:

1. Specialized occurrence was confirmed by keyness.
2. The cumulative frequency of headwords had to be higher than 30.
3. The word to be included had to be a keyword in a subcorpus.

As presented earlier, for the compilation of thematic wordlists related to the subcorpora specialized occurrence and keyness were the criteria of inclusion. Both the full TEC wordlist and the thematic word lists were verified with frequency and keyword outputs of *Range* programme. The first 30 items of the final wordlists can be seen in alphabetical order in *Table 10*. The full TEC word list is included in APPENDICES B and C, whereas the thematic word lists are given in APPENDICES D, E, F, G, H, and I.

Table 10: The first thirty items of the full TEC list and the thematic wordlists in alphabetical order

RANK	TEC WORDLIST	AIR TRAVEL	BEACH HOLIDAY	CITY TOURS	HOTELS ACCOMM.	RESTAURANT CATERING	WELLNESS SPA
		THEMATIC WORDLISTS					
	accommodation	aboard	abroad	architecture	accessible	ambience	acupressure
2.	airline	aircraft	affordable	boast	accommodate	appetiser	affordable
3.	ambience	airfare	ambience	cathedral	affordable	atmosphere	ambience
4.	amenities	airline	amenities	classical	ambience	attentive	amenities
5.	apartment	airport	anchor	cobbled	amenities	banquet	aqua gym
6.	appreciate	aisle	angling	cosmopolitan	apartment	bartender	aromatherapy
7.	archipelago	allowance	aquatic	court	architecture	beverages	atmosphere
8.	architecture	altitude	archipelago	crowded	arrival	bill	attentive
9.	area	amenities	bathe	cuisine	atmosphere	braised	balneotherapy
10.	arrival	aviation	beachgoer	culinary	attentive	brasserie	bathe
11.	atmosphere	baggage	breeze	depository	available	brew	bathrobe
12.	attraction	boarding	canoeing	display	baggage	brunch	belongings
13.	baggage	boarding pass	Caribbean	district	balcony	buffet	blissful
14.	beach	carry-on	climate	downtown	ballroom	champagne	bubble bath
15.	boast	check-in	convenience	dungeon	banquet	chilled	change room
16.	book	compartment	coral	eclectic	bartender	complain	chlorinated
17.	brasserie	connection	cottage	enchanting	bathrobe	course	client
18.	buffet	conveyor belt	crab	equestrian	bedding	crab	decor
19.	cancel	economy class	crew	escort	belongings	crispy	degree
20.	catering	emergency	cruise	ethnic	closet	cuisine	delightful
21.	chef	exceed	crystal	exhibition	concierge	culinary	discount
22.	climate	flight	cuisine	facade	cozy	cutlery	dissatisfied
23.	coast	flight attendant	culinary	facilities	cottage	delicacies	equipped
24.	compartment	gate	currents	fare	curtains	delicious	facilities
25.	complain	jet lag	deck	fortress	décor	dine	hammam
26.	concierge	limit	depth	fountain	disappointed	dish	healing
27.	convenience	lounge	dive	fresco	discount	flavourful	herbal
28.	cottage	luggage	dockyards	funicular	disturb	fried	hydro massage
29.	cruise	name tag	dolphin	gothic	dorm	gourmet	hygiene
30.	cuisine	overhead locker	drift	guide	dresser	gravy	impeccable

The scrutiny of the final word lists suggests that relatively low number of tokens provide sufficient dataset for the creation of pedagogically applicable vocabulary lists. Although these lists are not comprehensive, as they do not contain all the supposingly important key vocabulary

items, they can be fairly good starting points for the revision of known words, for eliciting familiar theme-specific words and expressions, and learning new ones.

4.6.7. The topic-related lexis of the subcorpora

It seemed to be reasonable to compile two separate listings of words based on their part-of-speech for the complete TEC vocabulary list, as the proportion of adjectives was high compared to nouns and verbs, and some items appeared both as a noun and verb in the texts. Appendix B enumerates the nouns (N) and verbs (V) of the integrated word list of the TEC, whereas in Appendix C the most frequent and typical descriptive and evaluative adjectives of tourism texts are shown, both lists comprise the words in alphabetical order. Some key nouns are members of nearly all the wordlists generated from the subcorpora, such as AMBIENCE, AMENITIES, ATMOSPHERE, CUISINE, FACILITIES, HOSPITALITY, OFFER, RECOMMEND, STAFF, or VIBE. The list of adjectives contains high frequency words that are also common in general English texts, including AWESOME, DELICIOUS or POPULAR, while some of them occur mainly as attributes of tourism entities such as IMPECCABLE, CULINARY or HOSPITABLE.

Some key words of the *Air travel* subcorpus have been presented in the pilot study, the final list of the most frequent and distinctive words related to aviation can be seen in APPENDIX D. The majority of words are highly technical; the frequency of occurrence of such words as AIRCRAFT, AISLE, ALTITUDE, RECLAIM, RUNWAY, TURBULENCE or SHUTTLE BUS is lower in general English texts. The proportion of semi-technical words is considerable; some of these words occur as keywords in other subcorpora as well, such as CHECK-IN or STAFF, while some key items are frequent in general English text too, BAGGAGE, PASSENGER, or PASSPORT, to illustrate. However, there are some deficiencies in the list probably due to the lower number of tokens (13, 582); synonyms of *flight attendant*, that is, words such as *air host/hostess*, or *steward/stewardess* were missing from the texts.

The vocabulary list of the *Beach holidays* subcorpus evokes the imagery of sea-sun-sand vacations; words such as DIVE, LAGOON, SUNBATHE or SNORKELLING recall pleasant memories, while others inspire to experience the vibe of beach holidays again, like BREEZE, ISLE, LIGHTHOUSE, PIER, WAVE or YACHT. Some of the key vocabulary is highly specific in this context; words like ANCHOR, ARCHIPELAGO, CURRENTS or PENINSULA less frequently occur in general English written or spoken texts, and attributes like CORAL,

CRYSTAL, MARINE or PRISTINE help envisage the waterscape of these holidays. The wordlist of the *Beach holidays* subcorpus is included in APPENDIX E.

The lexis of the *City tours* subcorpus encompasses mainly words denoting sights and attractions including CATHEDRAL, DUNGEON, FORTRESS, PARLIAMENT, TEMPLE or SANCTUARY; these buildings are described with attributes such as CLASSICAL, ECLECTIC, GOTHIC, HISTORICAL, or ROMANESQUE. The attractiveness of other places of interest involving FUNICULAR, FOUNTAINS, STATUES, SCULPTURES, EXHIBITIONS or VISTAS is augmented with adjectives like SPECTACULAR, STUNNING, PICTURESQUE, or ENCHANTING. The atmosphere of cities is made perceptible using attributes like COSMOPOLITAN, COBBLED, CROWDED, or VIBRANT. The persuasive-promotional role of lexical choices in texts related to city tours will be presented and discussed in Chapter 6. APPENDIX F comprises the core vocabulary of the *City tours* subcorpus.

The *Hotels and accommodation* subcorpus provided a fairly comprehensive listing of topic-related words that are shown in APPENDIX G. There are some recurring shared words with other subcorpora, including ACCOMMODATE, AFFORDABLE, AMBIENCE, AMENITIES, ATMOSPHERE, FACILITIES, HOSPITALITY, or IMPECCABLE. The number of technical words is relatively high; terms like CONCIERGE, FUNCTION ROOM, HOUSEKEEPER, HOSTEL, INN, MAID, ROOM RATE or SUITE occur mainly in this subfield of tourism. Comparing the words based on their part-of-speech the proportion of verbs is low, ACCOMMODATE, DISTURB, INCLUDE and REFUND are the most frequent key ones selected for the list, whereas the number of adjectives is higher, attributes like ACCESSIBLE, AFFORDABLE, ATTENTIVE, AVAILABLE, COZY, DISAPPOINTED, EQUIPPED, IMPECCABLE, LUXURIOUS, OVERLOOKING, OVERRATED, PLEASED, PRICEY, SPACIOUS and STYLISH reflect on one hand the descriptive persuasive style of tourism adverts, and the dissatisfaction in customer reviews on the other hand. Besides the aforementioned ones, numerous key nouns refer to hotel rooms; BALCONY, BEDDING, BATHROBE, CLOSET, CURTAINS, DRESSER, DUVET, FAUCET, MATTRESS, PILLOW, SHEET, SOCKET, TOILETRIES and WARDROBE are the most frequently appearing words in the subcorpus.

The word list of the *Restaurants and catering* subcorpus also comprises some words that are recurrent in other subcorpora, such as AMBIENCE, ATMOSPHERE, CUISINE, CULINARY, HOSPITALITY, IMPECCABLE or VIBE. The names of various ingredients frequently occur in

the texts including CRAB, LOBSTER, MUTTON, OYSTER, PRAWN/SHRIMP, SALMON, VENISON, or VEGGIES, as well as the names of different dishes like GRAVY, RATATOUILLE, SOUFFLÉ and STEW, desserts as SUNDAE and PASTRIES, or drinks like BEVERAGES, CHAMPAGNE, and SPIRITS. The technical vocabulary refers mainly to catering units as BRASSERIE, BUFFET or TAVERN, to different positions in the catering industry including BARTENDER, HOST/HOSTESS, or SOMMELIER, to utensils like CUTLERY, PLATTER, or NAPKIN, and to different cooking processes and techniques such as BRAISED, BREW, FRIED, MARINATED, POACHED, ROAST, SQUEEZED, STEAMED or STUFFED. Naturally, the adjectives describe mainly food and drink, besides the aforementioned ones CHILLED, CRISPY, DELICIOUS, FLAVOURFUL, GOURMET, MOUTH-WATERING, NOURISHING, ORGANIC, TASTEFUL/TASTELESS, and TENDER are the key attributes. The verbs in the texts are semi-technical, although they are suggestive of hospitality, they are often used in general English texts such as COMPLAIN, DINE, OFFER, ORDER, POUR or RECOMMEND. The vocabulary list of the *Restaurants and catering* subcorpus is included in APPENDIX H.

The terms of the *Wellness and spa* subcorpus are highly specific, words like ACUPRESSURE, AROMATHERAPY, AQUA GYM, BALNEOTHERAPY, HAMMAM, HYDRO MASSAGE, MUD BATH, PHYSIOTHERAPY, REFLEXOLOGY, and WHIRLPOOL quite rarely occur in general English context. There is group of key words that occur across the other subcorpora, such as AFFORDABLE, AMBIENCE, AMENITIES, ATMOSPHERE, FACILITIES, IMPECCABLE, VIBE and VOUCHER, while other words can be associated mainly with wellness and spa experiences, including BATHROBE, HYGIENE, MASSEUR/MASSEUSE, REGENERATION, SAUNA, SANITATION, TREATMENT, or WELLBEING. The verbs BATHE, INDULGE, PAMPER, REFRESH and UNWIND highlight the rewarding and beneficial features of spa experiences along with adjectives like BLISSFUL, DELIGHTFUL, HERBAL, HEALING, MEDICAL and RELIEVED. However, due to customer reviews, attributes with negative connotation like CHLORINATED, DISSATISFIED, NUDE, or OVERRATED are also present in the texts. The word list of the *Wellness and spa* subcorpus can be found in APPENDIX I.

4.6.8. Range, distribution and text coverage in the Tourism English Corpus
4.6.8.1. Text coverage of the corpus

Language learners' reading and listening comprehension skills are greatly affected by vocabulary knowledge. Schmitt, Jiang and Grabe (2011) found that there is a linear relationship between the percentage of vocabulary known in a text and the degree of reading comprehension;

their research confirmed that the larger the reader's vocabulary, the higher their understanding of the texts. As discussed in 4.5.4., earlier studies investigated this relationship to set a vocabulary knowledge threshold for successful reading comprehension; Laufer (1989) estimated that 95% percentage of known vocabulary is necessary for L2 learners to understand written texts, whereas Hu and Nation (2000) proposed that 98% lexical coverage is needed for independent reading without dictionary or glossary. Laufer and Ravenhorst-Kalovski (2010) supported Hu and Nation's (2000) findings; they claimed that 98% vocabulary knowledge is needful for independent reading, as 95% lexical knowledge would allow adequate reading comprehension when reading is assisted with some guidance. These studies shed light on the importance of vocabulary knowledge to reduce language learners' burden of reading comprehension, especially when they need academic or technical vocabulary to perform it successfully.

To interpret the text coverage of the *TEC*, it is worth having a look at the expected vocabulary size in order to attain 98% text coverage of different kinds of texts. *Table 11* shows the number of word families to achieve this coverage in various contexts.

Table 11: *Vocabulary sizes needed for attaining 98% text coverage of various kinds of texts that include proper nouns (adapted from Nation, 2006)*

Texts	98% coverage	Proper nouns
Spoken English	7,000 word families	1.3%
Children's movies	6,000 word families	1.5%
Newspapers	8,000 word families	5-6%
Novels	9,000 word families	1-2%

To examine the coverage that words at different frequency levels provide to the texts of a corpus, Nation (2006) developed frequency lists on the basis of the *BNC*, and investigated the coverage that they allow for in various spoken and written texts. He found that the first 1000 most frequent word families provide coverage of 78% to 81%, the second thousand families additional 8% to 9%, whereas the third thousand 3% to 5%, the fourth and fifth thousand 3%, from sixth to nine thousand 2% in addition, whereas from tenth to fourteenth thousand less than1%. Proper nouns account for 2% to 4% of written texts, and words that do not occur in the lists for 1% to 3%. It is estimated that it may be possible to achieve 98% coverage with the knowledge of 5000 words and proper nouns, yet teachers and learners should not rely on 5000-

word vocabulary knowledge, the larger the vocabulary, the easier to reach the appropriate coverage.

In Nation's (2007, p. 6) view if we want to work out how learners can boost their vocabulary size, their actual vocabulary size needs to be related to three main frequency levels; to high-frequency, mid-frequency, and low-frequency words. *Table 12* indicates the figures of word families at each level, and the learning procedures based at these levels, with reading comprehension in focus.

Table 12: *Main frequency levels of words related to word families and learning procedures (adapted from Nation, 2007)*

Level	1000-word family lists	Learning procedures
High frequency	1000-2000	Reading graded readers
		Deliberate teaching and learning
Mid-frequency	3000-9000	Reading mid-frequency readers
		Deliberate learning
Low frequency	10,000 onward	Wide reading
		Specialised study of a subject area

The unit of the investigation on text coverage is the word family, based on the observation that knowing one member of the word family and word-building processes with using affixes help work out the meaning of the unknown members of the family. However, there is a substantial difference in the number of words based on the unit of counts; the first 1000 word families contain 6,857 word types, thus each word family has nearly seven words per family on average, however, there is a decrease in the number of types per family at the mid-frequency and low-frequency levels.

The coverage of the TEC was analysed using a new version of the CLT vocabulary profiler that matches the texts to frequency lists created on the basis of the BNC. The output of the *Vocabprofiler* shows what percentage of the texts is covered at different levels, ranging from the most frequent 1000-word (K1) level to 20,000 (20K) level. Words that are not in the most frequent 20K levels are marked as 'off-list' in the output. Proper nouns that are personal or geographic names (e.g., Michael, New York) appear in the first thousand most frequent words, assuming that they do not belong to a particular language, thus problems with comprehension

cannot be attributed to them. However, proper nouns made up of regular words (e.g., Golden Gate Bridge, Discourse Analysis) appear at their respective level, and some rare proper nouns occur as 'off-list' words. *Table 13* presents the coverage of the TEC texts.

Table 13*: Coverage of the TEC texts based on BNC frequency lists, generated by CLT Vocabprofiler*

Level	Tokens	Coverage %	Average cumulative coverage
K1	140,242	77.660	77.660
K2	17,991	9.963	87.623
K3	6,856	3.797	91.420
K4	3,134	1.735	93.155
K5	3,535	1.958	95.113
Coverage 95%			
K6	1,606	0.889	96.002
K7	1,050	0.581	96.583
K8	652	0.361	96.944
K9	457	0.253	97.197
K10-K20	1,472	0.816	98.013
Coverage 98%			
Off-list	3525	1.954	99.967

These figures largely correspond with Nation's (2006) findings on the coverage of texts based on frequency lists generated from the BNC; the first 1000 words provide nearly 78% coverage, the second thousand additional 9.9%, the third thousand 3.7%, whereas the fourth and fifth thousand words provide less than 2% coverage in addition, compared to the 3% cumulation of coverage that Nation (2006) observed at this level. These percentages of vocabulary coverage suggest that tourism language is made up of general words, sub-technical vocabulary, and tourism terms. In comparison with other specialised fields, Nation (2008, p. 36) found that the vocabulary of anatomy texts, applying the frequency approach, had as low as 64.6 % coverage on the basis of general standard lists, implying that their lexis is highly technical.

To achieve 95% coverage in the comprehension of written tourism texts learners need to be familiar with the first 5000 words; this figure suggests that the students might not be able to cope with the inference of unknown words without assistance. A learner with a 2000-word

vocabulary will be familiar to 87.6 percent of the words of the TEC texts; supposingly, they need at least a 3000 word-family vocabulary to achieve reasonable comprehension, and, according to Nation (2006) that size of vocabulary is needed for deliberate learning. Thus, besides dictionaries and glossaries, explicit vocabulary teaching and exposure to unfamiliar words can enhance their reading comprehension skills.

To support the findings on text coverage *Range* programme can be applied to compare the TEC texts against its base vocabulary lists; the most frequent 1000, the second 1000 most frequent words of English, and base word list three that includes the most frequent words from a wide range of texts of secondary school and university subjects. In *Table 14* the text coverage of TEC is presented, framed up on *Range* base word lists.

Table 14: *Text coverage of the TEC framed up on Range programme base word lists*

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	134,010/72.1	2,607/20.8	969
two	16,337/ 8.8	1,522/12.1	740
three	6,341/ 3.4	896/ 7.1	434
not in the lists	29,127/15.7	7,519/59.9	n. a.
Total:	185,815	12,544	2,143

All the words in the base lists are counted as members of word families, whereas the remaining words are counted as types. The figures show that 72.1% of the running words of the TEC are in the first 1000 base list; this coverage is somewhat lower than the 77.66 % coverage rendered by the *CLT Vocabprofiler* on the basis of BNC frequency lists. The second 1000 base word list also yielded lower coverage, 8.8%, compared with the 9.963% coverage provided by the *Vocabprofiler*. However, adding up these figures with the 3.5 % coverage of the third base list, the total coverage of the TEC texts is only 84.3% that does not allow for successful comprehension, indicating the need for assistance and explicit vocabulary learning.

4.6.8.2. Range and distribution of some key terms in the corpus

Applying the concordance plot function of *Antconc 3.5.8* the dispersion of a target word is made detectable in the corpus. This function can help to identify whether a word appears only in a section of the corpus due to its high topic specificity, or it is evenly distributed. Even

distribution of a word across the corpus indicates that it is less likely to be topic-related or specific in a particular subcorpus. The subsequent position of subcorpora that are included in alphabetical order help localize the distribution of target headwords, namely, the *air travel, beach holidays, city tours, hotels and accommodation, restaurants and catering*, and *wellness and spa* holidays subcorpora. As *Figure 4* demonstrates the word *airline* is highly subject specific, its occurrence is restricted to the *Air travel* subcorpus of the TEC.
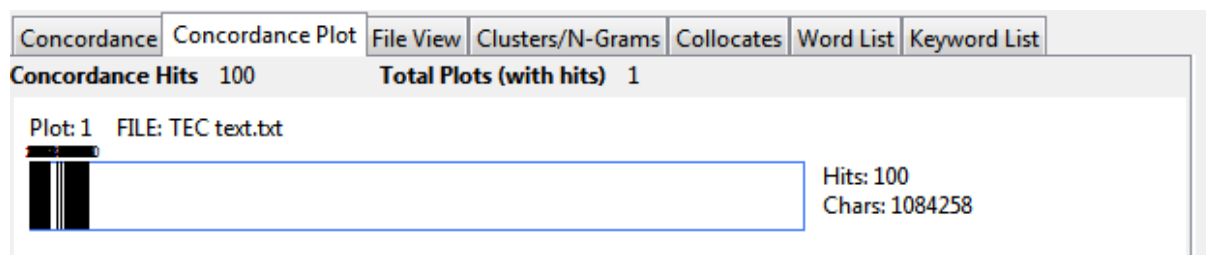
Figure 4: Distribution of the headword 'airline' in the Tourism English Corpus

Figure 5 displays the range and distribution of the search term *beach* in the TEC. It can be detected that '*beach*' has low occurrence in the *Air travel* subcorpus (2), its frequency of occurrence is obviously the highest in the *Beach holidays* subcorpus (583), but this key word is infrequent in the *City tours* subcorpus (14). It appears quite frequently in the *Hotels and accommodation* section (113), its frequency of occurrence is low in the *Restaurants and catering* segment (8), and it has limited occurrences in the *Wellness and spa* subcorpus (5). The frequency of occurrence figures is displayed in *Table 7*.
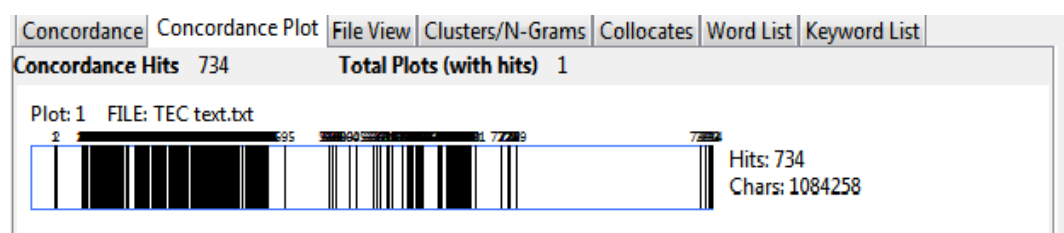
Figure 5: Distribution of the headword 'beach' in the Tourism English Corpus

Having a look at the distribution of search terms that appeared in the wordlists of several subcorpora, it is apparent that some of them have a wide range, like *affordable or facilities*, as they occur in nearly all the subcorpora, and their distribution is fairly even in the texts of the TEC. However, other common shared words of the vocabulary lists including *staff*, *atmosphere* or *amenities* have a narrow range as they occur mainly in those subcorpora where their presence is more relevant.

4.7. Pedagogical implications of wordlists

4.7.1. Focus on teaching and learning technical vocabulary

Learners of different tourism professions need to understand not only specific tourism terms, but general words that denote a different meaning in tourism context. Some lexical items that have a well-known meaning and are frequently used in general English may have another meaning that is used exclusively in tourism context; therefore, they are likely to be problematic for language learners. A collection of these words can help learners recognise the differences between their uses in general and tourism English texts. *Table 15* presents some examples from the word lists whose meaning differs in general and tourism English context. The definitions are extracted from the *Oxford Advanced Learner Dictionary*.

Table 15: *Examples of high frequency general words with technical meaning*

WORD	GENERAL MEANING	MEANING IN TOURISM
lounge	synonym of living room	public room in a hotel and at an airport
gate	a door-like barrier in a fence or hedge	a way out of an airport building to an aircraft
transfer	move from one place to another	change vehicle during a journey
connection	synonym of link, the act or state of being connected	a train, bus, or plane that a passenger can take soon after getting off another
attraction	a feeling of liking somebody	an interesting place to go or thing to do

Another means of enhancing technical vocabulary is the compilation of annotated vocabulary lists that include the POS, definition, and collocations of words, as well as sample sentences on their usage in the texts of the corpus. Annotated lists allow learners to study how a particular word is used in the target corpus, this way making the process of vocabulary learning more effectual. The annotated lists focus on the contextual meanings of words and their usage in a corpus, the presentation of their part of speech, their frequent collocation patterns and definitions, additionally, sample sentences help clarify meaning and promote retention. *Table 16* shows an annotated vocabulary list excerpt developed from the *Hotels and accommodation* subcorpus.

Table 16: Example of annotated tourism English word list

WORD	PART-OF-SPEECH	DEFINITION	COLLOCATIONS	SAMPLE SENTENCE
ambience	noun	the character and atmosphere of a place	ambience of a hotel	We are proud to *offer the ambience of the hotel* with continuously renovated rooms.
amenities	noun (plural)	a feature or service that makes a place pleasant and comfortable	access to amenities; upscale, modern, exceptional amenities	Our four-diamond hotel rooms offer attentive service, *upscale amenities* and blissful tranquillity in a sensual setting.
concierge	noun	a person in a hotel who helps guests by giving them information, arranging theatre tickets, etc.	concierge service concierge level	The amenities include 8 conference halls, a business centre, *a concierge service*, and art gallery.
cozy	adjective	warm, comfortable and safe	cozy guest room cozy atmosphere	The hotel offers 229 *cozy* and spacious *guest rooms*.
disturb	verb	to interrupt somebody when they are trying to work, sleep, etc.	do not disturb sign	For us using the do not disturb sign is a style statement, not a declaration.
dorm	noun	-a building for university or college students to live in - a room for several people to sleep in	dorm accommodation dorm room affordable dorm	Most hostels with dorm accommodation also offer private rooms at affordable rate.
duvet	noun	a large cloth bag filled with feathers or other soft material on the top of a bed to keep yourself warm	nice and fluffy duvet sheet and duvet	The duvet is nice and fluffy but warm.

The definitions were taken from the *Oxford Advanced Learner Dictionary*, since they are written for language learners, they follow an easily understandable pattern. The sample sentences had to be shortened and simplified to make them easier to understand, however, they aimed to include recurring high frequency terms to make the meaning of the target word clear, and to create and deepen associative links in the mental lexicon.

4.7.2. Exploitation of the corpus in vocabulary development

A wide range of vocabulary teaching and learning activities based on corpora can be applied in tourism classes that students enjoy and appreciate, in my teaching practice the printouts of multiple-matching and gap-filling activities are the most popular ones with students. Corpus-based reading comprehension tasks, matching words with their definitions, finding synonyms and antonyms or the most typical collocations of a certain word facilitate access to the lexicogrammatical structures of tourism language. It is rewarding to familiarize students with the use of online concordancers, dictionaries or thesauri, they are particularly helpful when learners come across new vocabulary items or do assignments.

The texts and wordlists of the TEC can be exploited in various ways to extend and deepen tourism students' word knowledge. Eliciting familiar words with the help of vocabulary lists and learning new terms as lead-in activities can enhance students' motivation and engagement in performing the subsequent tasks. The word lists and the concordance lines of the corpus can be exploited to teach synonymous nouns, adjectives and verbs; to illustrate, *ambience* and *vibe* frequently occur as synonyms of '*atmosphere*', various alternatives of '*expensive*' as *overrated*, *overpriced,* and *pricey* appear in the texts, or '*accommodate*' can be replaced with *lodge* or *put up*. Observing the attributes of keywords in concordance lines is also a good means of learning synonymous adjectives, as an instance, a '*fine dish*' can be described as *flavourful*, *mouth-watering*, *tasteful*, *delicious*, or simply *yummy*.

The frequency and keyword lists provide ample examples for teaching and learning words that differ either in word form or in spelling in British and American English. A bunch of words used in American English vs British English can be detected in the frequency and keyword lists and in the thematic word lists as well, such as *check-bill*, *elevator-lift*, *faucet-tap*, *closet-cupboard*, *subway-underground*, *streetcar-tram*, *shrimp-prawn*, or *cookie-biscuit*, just to mention a few. Spelling differences between British and American English can also be highlighted; words like *cosy/cozy*, *flavour/flavor*, or *harbour/harbor* occur as different items in the frequency and keyword lists.

4.7.3. Applying the tools of the Compleat Lexical Tutor in vocabulary teaching

A growing number of online resources are available to help students increase their vocabulary on the basis of wordlists, either for individual vocabulary learning or for classroom activities. One of the major widely applicable resources is Cobb's remarkable *Compleat Lexical Tutor* (CLT, https://www.lextutor.ca/) that has various tools for analysing texts and supporting vocabulary learning, its functionalities are great aids for teachers to create motivating and exciting corpus-based vocabulary teaching and testing tasks and activities. It helps generate exercises based on the examples of ready-made or self-compiled corpora to make classes more interactive and interesting, and provides self-access learning opportunities for students. Learners can test their vocabulary and vocabulary levels, compare their active and passive vocabulary, and acquire new words. *Concordancers* not only allows the observation of target words in concordance lines, but also helps create gap-filling activities, the *MultiConc* interactive *Quiz Option* can be used either in pen-and-paper or in interactive version. With the

help of the *Concordancers* function students can deduce rules on grammar and on vocabulary use, examine near-synonyms, count recurrences, and as an instance, find out whether 'cheap hotels' or 'expensive hotels' are more common in tourism texts.

To test students' vocabulary or create interesting customized tasks for practice targeted cloze-tests, multiple concordance or multiple-choice quizzes can be compiled with the help of the exercise builders in the *Teachers* section. Applying the *Clickers Builder* functionality multiple choice vocabulary quizzes can be created, students can send their responses from mobile phone or computer, thus, teachers come to know quickly what learners know and what needs further elaboration. Another tutorial tool of CLT that promotes vocabulary development is the *Flashcards Builder* programme that assembles up to 100 paired associates including word and its definition, synonyms or antonyms. These activities help practice the target vocabulary so that learners could cope more effectively with their reading comprehension and professional communicative tasks, as well as enhance students' learning strategies. The *Vocab Test* tool provides mainly frequency-level based multiple choice tests for the self-assessment of receptive and productive vocabulary, and offers the option for teachers to compile their own tests including yes-or-no mobile phone based tests.

4.8. Summary

This chapter presented some ways how the principles and methods of corpus linguistics can be applied in corpus compilation and vocabulary list creation for pedagogical purposes. The *Tourism English Corpus* was developed to help learners acquire the most important lexis of tourism with the key words assembled in thematic vocabulary lists and in an integrated TEC word list, as well as to gain higher coverage of tourism texts. The findings of the study can be summarized as follows:

- Although the word lists have some shortcomings such as the lack of some important keywords, not treating homoforms, especially homonyms as different items, and not identifying separate members as compound nouns, it has been confirmed that the word lists can encompass distinct sets of vocabulary related to subfields of tourism. As the number of shared items across the thematic word lists is limited, words associated with that particular domain can be selected for topic-based vocabulary teaching.

- The analysis of text coverage and the profile of the TEC revealed that the texts contain a relatively large proportion of words from the general standard list of the BNC, as the text coverage, in terms of the first two thousand most frequent words, was nearly 88%. It indicates that a large proportion of the vocabulary of tourism texts, hence the items of the word lists, are not highly technical but fall in the categories that are considered as general and sub-technical vocabulary.

- Applying annotated frequency lists of tourism terms can reduce study time, allowing focusing on the core technical vocabulary and on the usage of these words in specific context.

- In addition to serving as a basis for lexical syllabi, the thematic word lists can be utilized in creating contextualised teaching materials and vocabulary learning activities.

- Although this size of a corpus proved to be adequate for generating fairly applicable word lists, further extension of the corpus would yield more comprehensive vocabulary lists, especially in terms of low frequency items. Another line of further analysis could be the categorization of keywords based on the level of their difficulty.

Besides contextualised vocabulary learning, word-focused lists can orientate vocabulary learning activities that help learners boost their receptive and productive vocabulary. As the aim of vocabulary learning activities is to achieve long-term retention, word lists can be used effectively to increase the amount of acquired words and the rate of vocabulary extension, since these lists present the most frequently occurring and most significant vocabulary of a specific language or field. Word frequency has various effects on vocabulary acquisition; retention is positively affected by frequently occurring words as learners are subjects to increased exposures. Likewise, lexical access is beneficially influenced by high frequency words due to a phenomenon called word frequency effect, that is, the recognition time of frequently seen words is faster than that of less frequently occurring ones (Segui et al., 1982). Since vocabulary lists gather words that are likely to be unknown and more time-consuming to acquire by reading, they are practical guidelines for language teachers and ESP practitioners when compiling course material, to ensure that frequently used words are incorporated in the syllabus or lexical syllabi.

The Tourism English Corpus has other potentials to exploit, in the following chapter it allows for the investigations on the persuasive and promotional discourse functions of tourism language, whereas in Chapter 7 for the examination of lexical bundles and collocations in tourism texts.

CHAPTER FIVE

THE PROMOTIONAL AND PERSUASIVE DISCOURSE FUNCTIONS IN THE LANGUAGE OF TOURISM

A CORPUS-DRIVEN ANALYSIS OF TRAVEL ARTICLES

5.1. Introduction

As part of the ever-increasing research on the special language and discourse of tourism, the study in this chapter inquires into how the persuasive discourse function, an effective communicative means of attracting potential customers, occurs in the texts of a small-scale tourism corpus. The study focuses on the investigation of lexical and grammatical choices in travel articles in the *city tours subcorpus*, to examine in what ways the selection and recurrence of certain lexical and syntactic patterns manifest and demonstrate the persuasive promotional function of this special discourse. Following an account of the theoretical framework, the results of a corpus-driven analysis are presented concerning the use of keywords, attributive clusters, intensifiers, imperatives, and ego-targeting person pronouns, as markers of persuasion in the articles related to city tours. The results suggest that further research is needed on the discourse functions of the different registers of tourism language, as the classroom application of the findings can contribute to the development of tourism students' discursive competence, consequently, to the enhancement of their professional communicative competence.

5.2. Reasons for investigating the discourse of tourism

Tourism has been one of the most flourishing industries worldwide (Scowsill, 2017, p. 3) that has had a great impact on the economy of nations and geographical areas. Tourism, as it has been defined by the *World Tourism Organisation* (2014), is 'a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes'. Due to the previous growth of the tourism sector, the considerable number of tourists and overseas holidays, tourism as a multidiscipline including promotional marketing and travel journalism has created its special language that has become the focus of numerous studies. The promotion of tourism products

brought forth the specific tourism discourse, in which the language is used as a medium to persuade visitors, and the means of verbal and visual communication such as language, images and sounds are strategically combined to inform and attract potential travellers. Promotional media, especially advertising and journalism, apply multimodal visual and communicative strategies to inform and influence the target audience in decision-making; besides creating an appealing imagery of destinations and services with visual aids, the creative and persuasive use of language increases their attractiveness.

Persuasion is an essential element of tourism discourse as it may enhance the truth value of the information on destinations and services, and, additionally, it induces a 'worth-visiting' image in prospective tourists. So far, studies on the persuasive function of tourism discourse have been conducted concerning only promotional media such as tourism advertisements, brochures, and web-pages (Hassan et. al, 2008; Ling, p, 2008; Luo & Huang, 2015; Manca, 2019; Mocini, 2005; Sona, 2015; Sparks et. al, 2013). However, several aspects of this special discourse remain unexplored, especially in terms of teaching this specific language and its discursive features to would-be tourism personnel. Corpus-driven research on the language and discourse of tourism is based mainly on self-compiled corpora, as large-scale tourism corpora are publicly not available. This empirical study intends to add some specific details to the line of corpus-based research on the persuasive discourse function of tourism language from a different perspective, investigating in what ways persuasiveness is reflected in the choices of lexical and syntactic structures in a specific register, travel articles.

5.3. The language of tourism

Since promotional media with the creative, descriptive, and persuasive use of tourism language evokes the imagery and fascination of travel to affect potential clients' decisions on holiday-making, this special language reflects the commercial, marketing aspect of the genre. The language of tourism depicts the tourist destinations and services in persuasive, encouraging and alluring manner, and makes them visual with dynamic images and vivid colours (Francesconi, 2014). Hence, the texts of brochures, travel articles, and advertisements are crammed with descriptive and evaluative adjectives, superlatives and intensifiers to influence and even manipulate the attitude and behaviour of prospective tourists.

Tourism promotion is an essential means of providing information for the potential tourists, correspondingly, the promotional-commercial discourse of tourism takes place in a business-consumer context. With the help of the Internet the tourism sector has found a rapid and effective way to reach the target group of potential customers; the articles of online tourism magazines, commercial and official websites, the web pages of national and local tourism boards attempt to allure and encourage potential visitors to become actual tourists. Social media as a promotional tool allows tourism operators to make contact directly with potential travellers; it is widely accepted as an influential factor of clients 'decision making and in the evaluation of the destination (Pierini, 2009, p. 97). Nevertheless, it was revealed in a representative survey conducted in 2018 that in Hungarian tourists' decision making the Internet does not have a predominant role; although they obtain their preliminary information on destinations from Internet sources, the opinions of friends and acquaintances and information provided in travel catalogues are also important inspiring factors (Gonda & Csapó, 2019, p. 109).

Tourism has a discourse of its own in which promotion is manifested in highly persuasive manner. Tourism Web sites are the main platforms of the genre, the travel articles and web adverts contain informational-promotional and commercial sections, sometimes introducing stereotypical features of the promoted entity. Their communicative purpose is to convince the target audience of a market segment, either with attractive images or linguistic devices. Although tourism discourse reflects different levels of specialisation, the main aim is to reach a wide target audience who are non-specialists of the field, thus promotional discourse has a low level of specialisation and tends to resemble to general language (Pierini, 2009, pp. 97-98).

Accordingly, the technical vocabulary of English in tourism consists of general words drawn from everyday language, semi- or sub-technical words shared with other disciplines, and technical terms that are used only in tourism context referring to specific concepts, like *tourism ecosystem* or *tourism yield*. The most widely used vocabulary in English for tourism, also known as core business vocabulary, is sub-technical; however, advanced learners and specialists tend to use highly technical terms and collocations (Brunet, 2011, p. 154). Besides general words, technical and sub-technical terms, a large proportion of proper names, words and idioms borrowed from the languages of destinations characterise tourism English (Wang, 2010, pp. 337-338). Corpus-driven analysis of loan words and expressions in tourism language disclosed that foreign terms can be considered a pragmatic marker of social attitude in tourism discourse, and reveal deeper intercultural relations of destinations (Gandin, 2014, pp. 2-14).

As the fundamental element of tourism discourse is the presentation of the tourism product, its characteristics are described either explicitly or implicitly with positive terms, this way descriptions and evaluations are intertwined in tourism texts. Two basic strategies are applied for description or evaluation; the positive features of the product or service are emphasised with highly evaluative vocabulary, and, with the help of lexico-grammatical structures and linguistic devices the readers' emotions are affected to motivate them to purchase the product. The descriptions and evaluations reflect the knowledgeable tourism experts' viewpoints who make judgements on the tourism entity; thus, it is a central task in the investigation of tourism discourse to explore how the means of vocabulary selection help convey their viewpoint to the encoders, and have an impact on their perception of that entity (Pierini, 2009, p. 98).

To highlight the values of the tourism product and to affect the consumers 'emotions, tourism genres make great use of comparisons, metaphors and similes, keywords and keying, languaging, that is, impressive use of foreign words, and omission (Dann, 1996, pp. 37-39). Comparisons with familiar places and experiences make the tourists more acquainted with novelty, the use of keywords aims to boost the imagination of tourists, and real or invented foreign words induce the feeling of inferiority in the tourists who depend on the guiding of the message (Nigro, 2006, p. 188). However, the language of tourism is not a well-established linguistic domain as it is a combination of different disciplines and technical fields such arts, history, geography, transportation or marketing (Calvi, 2005, p. 43 quoted in Nigro, 2006, pp. 188-189). On the other hand, tourism is a social fact that is connected with other subsystems such as economy and politics (Cordero, 2011), thus, this peculiar language is made up of a wide range of lexical, stylistic and pragmatic features that are intertwined with different specialised languages and registers (Gandin, 2013, p. 326).

5.3. 1. Tourism discourse

Discourse, a term that applies to both written and spoken language, according to Stubbs's (1983, p. 1) widely known definition is the language 'above the sentence or above the clause'. Opposed to this view, Sharma (2010) regards discourse as any series of speech events or any combination of sentences both above and below the sentence level, wherein successive sentences, phrases, or utterances hang together. Gee (2011) regards the texts and utterances of a discourse as communicative units that are embedded in social and cultural practices, implying how people manage to interact, how they communicate within particular groups and with different cultures. Thus, discourse is shaped by the world and by the language, as well as shaping the world and

the language. Discourse is also influenced by the medium in which it occurs, and the purpose of the text also has an impact on the text. Gee (2011) argues that discourse includes different styles of languages and involves different socially shaped identities, particular ways of thinking, believing, valuing and knowing. In his (2005) interpretation: 'Discourse is a 'dance' that exists in the abstract as a coordinated pattern of words, deeds, values, beliefs, symbols, tools, objects, times, and places in the here and now as a performance... (p.19) '. Therefore, discourses, as performances of different identities are socially constructed. Thereby, within the frame of discourse analysis and sociolinguistics, a line of research has been directed towards examining the language of tourism from the perspectives of touristic experience, identity, performance and authenticity (Fox, 2008, p. 14).

The specific tourism discourse is explained by Gotti (2003, p. 24) as 'the specialist use of the language in contexts which are typical of a specialized community stretching across the academic, the professional, and the occupational areas of knowledge and practice'. The promotional and marketing aspects of tourism are reflected in Dann's (1996, p. 2) interpretation of this domain-specific discourse; 'tourism … has a discourse of its own. …the language of tourism attempts to persuade, lure, woo and seduce millions of human beings and, in doing so, convert them from potential into actual clients'. Yet, prior to Dann's research in the 1990s, as he remarked, 'no one has comprehensively analysed this language as a phenomenon in its own right. Certainly, there have been some studies which have alluded to the linguistic features of tourism promotion, but none so far brought them together and systematically examined tourism as a language per se' (p. 2).

As the first sociolinguistic analysis of language and tourism was carried out by Graham Dann, his by now classic book, The *Language of Tourism: A Sociolinguistic Perspective* (1996) designated a direction for further sociolinguistic explorations of tourism language. Assuming, that tourism has a discourse of its own and it fulfils numerous social functions, Dann created a novel understanding of the language of tourism. Tourism discourse can function:

- o as an instrument of customers' active involvement in the process of consumption and in the co-creation of the language that inspires them to consume,
- o as a means of social control implying norms and values,
- o and as a medium of socialisation employed in the development of identities.

The adoption of sociolinguistics to support the theory of tourism gives new insights into the complex and intricate relationship between language and tourism (Fox, 2008, p. 20). In the

discourse of tourism language use is part of a social discursive practice, beyond the literal meaning tourists infer and grasp the intended convincing and influential meaning as well.

Since Dann's (1996) landmark study several features of tourism language have been investigated from different perspectives including sociolinguistic, linguistic, cultural, pragmatic, discourse, semantic, and corpus-based ones (Alcantar, 2007; Capelli, 2006; Castello, 2013; Francesconi, 2012, 2014; Gotti, 2006, 2011; Jaworsky & Thurlow, 2010; Maci, 2012; Nigro, 2006; Thurlow & Heller, 2014; Tognini-Bonelli & Manca, 2002; Wang, 2010). The investigation of sociolinguistic, discourse and lexical features of tourism language has gradually become prominent in researching the interrelation of language and tourism (Dann, 1996; Francesconi, 2014; Gotti; 2006, Maci, 2010; 2012; Nigro, 2006; Wang, 2010).

Castello (2002) and Lam (2004) are considered the initiators of corpus-based tourism text analyses; since the compilation of their large-scale corpora and milestone studies remarkable undertakings have been carried out in corpus-driven tourism language research (Chujo, et. al 2006; Kang 2011; Nigro 2006). Gotti (2006) emphasized the importance of pragmatic communicative function in determining a tourism text's genre, whereas Gandin (2014) examined whether loan words and expressions can be considered pragmatic markers of social attitude in this distinctive discourse. Francesconi (2014) searched for the means of integrating applied linguistics with multimodal discourse analysis, shifting the textual approach of tourism discourse to semiotic perspective.

Grounded on Dann's (1996) work Luo and Huang (2015), as well as Nigro (2006) carried out research on the special terminology and discourse of guidebooks, whereas Rosypalova (2012) conducted a corpus-based analysis on the written and spoken discourse of tour guides. However, few studies have investigated promotion and persuasion from discourse analytic perspective (Edwards & Curado, 2003; Hassan et. al, 2008; Manca, 2019). Further research that employs multimodal discourse analysis approach in the investigation of tourism websites could be a significant move forward in this area.

5.3.2. The communicative and persuasive functions in tourism discourse

In Nigro's view (2006, p. 189) the language of tourism fulfils three communicative functions; *vocative* (promotional), *expressive* and *referential*. The main *vocative function* of promotional

materials such as brochures, leaflets, and advertisements is to attract the attention of potential customers, whereas travelogues and tourism journals have a predominant *expressive function* as they focus on individual tourists' accounts of travel experiences. The *descriptive referential function* is predominant in guidebooks or descriptive panels of museums that present tourist attractions with technical terms and provide objective information. However, as Nigro (2006, p. 189) points out, there is no clear-cut distinction of linguistic functions in tourism texts as they generally appear simultaneously.

Persuasiveness is a crucial element of the vocative and expressive communicative functions in promotional materials and travel journalism, as the alluring, personalised, and detailed descriptions of the destinations make the information credible for potential clients. Besides attracting the hope-for tourists' attention, the application of persuasive discourse techniques helps them to refine their image towards the destination by highlighting its specialities, introducing favourable details, and offering incentives (Luo & Huang, 2015, p. 8). An important and unique persuasive device of tourism discourse is tautology (Dann, 1996, p. 66), that is, tourists take the content of tourism texts for granted, and information that sounds plausible has a positive impact on consumers' beliefs and decisions.

Tautology is a controversial issue in travel journalism. Opposed to the most important textual feature of journalism, wherein the content counts as true, travel articles are inherently biased and may contain fictional elements as well. In addition, travel journalism is increasingly conducted online by travel journalists and freelancers who are not experts in the field, which definitely has a negative impact on the content of the articles (Hanus & Fürsich, 2014, pp. 7-8). To compensate these deficiencies in the credibility and trustworthiness of the content, the use of persuasive discursive devices is central in travel journalism.

Besides travel journalists, tourism personnel and ESP students as would-be tourism professionals also need to be aware of the distinctive stylistic and discursive features of tourism language, in which persuasiveness is an indispensable element. With the ultimate aim that the findings can be implemented in teaching the characteristic discursive features and lexico-grammatical patterns of this specialised language, I sought for different means and techniques of persuasion in travel articles. The study investigates in what ways lexico-grammatical patterns such as keywords, attributive clusters, intensifiers, adverbials, pronouns, and imperatives occur as persuasive devices to increase plausibility and trustworthiness in this specific register.

5.4. Research questions

Adopting Halliday's (1991, p. 32) integrating approach of lexis and grammar the research focused on the following questions:

1. What lexico-grammatical persuasive discursive tools can be detected in a corpus comprising a scant amount of running words?
2. In what ways is the persuasive discourse function of tourism texts manifested in the lexical choices of the articles?
3. In what ways do grammatical choices enhance persuasiveness in tourism texts?

The exploration and identification of the persuasive discourse function in syntactic structures seemed to be the most intricate and perplexing. Since word selection in tourism texts highly depends on the communicative goal that they intend to achieve, the lexical analysis of the texts cannot be separated from certain syntactic patterns that help convey the intention to persuade, beyond the literal meaning. As the lexico-grammatical choices hinge on and intertwine with the intended communicative goals, the lexical and grammatical aspects cannot be examined separately, however, the analysis is organised and moves along the continuum of lexis, syntax, and discourse.

5.5. Methodology

Several methodologies and models are applied to the identification of promotional and persuasive strategies that occur in travel articles; Halliday's lexico-grammar approach, Kress and van Leeuwen's multimodal framework of communication, Cook's schemata model in discourse processing, Dann's concept of tourism genres, and the CL approach and methodologies.

5.6. Corpus compilation and data analysis

Following Dann's (1996) concept that different registers of tourism language are representative of different domains in the tourism industry, I compiled a corpus of travel articles comprising 30,846 tokens on the topic of 'city tours' that was integrated into the TEC as a subcorpus. The articles and relevant parts of articles were mainly selected from a free website that has altered its content, *Articlecity,* http://www.articlecity.com/search.shtml. Relevant parts of articles were

also selected from the ones sent to my e-mail address by *Google Alerts* on the keywords *city tours*, and some extracts of additional articles were sorted out from different tourism websites with the help of *WebCorp*, *www.webcorp.org.uk*, an online concordancer that lists articles on the required key words. The texts were saved in plain text format, and had to be cleaned from the disturbing 'noise'; the unwanted advertisements and contact details were removed. The subcorpus, dubbed '*City tours*', is available online at http://bit.ly/GH6VnA.

Data processing was administered with Anthony's versatile freeware, *AntConc3.2.4.w*, and Cobb's *Compleat Lexical Tutor* (CLT). To generate concordance lines, analyze lexical bundles and word frequencies I used mostly the *AntConc* Concordance and the Clusters (N-Grams) tool that scans the entire corpus for 'N' (1, 2 or 3 words) length clusters. First, I generated frequency and keyword lists, then investigated the occurrence of attributes including superlative forms with a wildcard (*), and also looked at frequent attribute-noun clusters including adjectival sequences and comparisons. Afterwards, I examined the persuasive role of intensifiers and the epistemic adverbial 'certainly', the next step was the investigation on the infringement of taboo, imperatives and exclamations, and finally, the ego-targeting use of person pronouns was looked into.

5.7. Results and discussion

5.7.1. Frequency

The lexis of the articles demonstrates a distinctive feature of tourism discourse, the most frequent words and keywords tend to have positive connotation; they convey the implicit message to the readers that they would have a wonderful time at the destination, full of exciting activities. Frequently occurring words that evoke positive associations are effective means of persuasion as they boost the imagination of tourists, meeting their expectations and desire to escape from everyday life by creating the image of a pleasant experience. The most frequent common nouns, adjectives, and main verbs can be seen in *Table 17*.

Table 17: *Frequency (FQ) lists of common nouns, adjectives and main verbs*

Rank	Common noun	FQ	Adjective	FQ	Main verb	FQ
1	city/cities	469	beautiful	74	offer/s	84
2	tour/s	467	great	56	see	83
3	world	87	new	46	visit	81
4	history	70	popular	45	enjoy	60
5	place	68	famous	42	like	54
6	travel	66	old	32	said	51
7	day	58	wonderful	30	get	50
8	attraction/s	57	historic	31	take	46
9	time	52	unique	26	explore	41
10	people	46	top	28	make	36

Besides the thought-provoking nouns, positive meaning can be associated to all the adjectives, in addition, *unique* and *top* refer to the exceptional and superb qualities of tourist attractions. The most frequent verb, *offer*, is a hidden reference to influence tourists to take part in sightseeing tours, and other frequent verbs like *see*, *visit, enjoy* and *explore* also encourage them not to miss this pleasurable and attractive experience.

5.7.2. Keywords and keying

Keywords are 'indicators of texts or domain specific terms' (Bowker & Pearson, 2002, 149), and, with their repetitiveness and salience, they are effective means of drawing people's attention in tourism texts. Keyness, a quality of words or lexical chunks, is text-dependent; keyness suggests that 'the wordings are prominent in some way' in that particular text (Scott, 2010, p. 44). The prominence and outstandingness of keywords in tourism texts lead to a general perception of the content, it may help perceive the style and create an imagery of the journey. Dann (1996, pp. 174-175) defined the property of keywords and keying 'as a series of attributes of a destination ... which corresponds to the requirements of potential tourists using 'appropriate language able to give an aura of genuineness and authenticity to the destination promoted '.

To generate a keyword list the CLT's keyword function was used as it determines keyness related to a larger reference corpus, the *British National Corpus* (BNC). As *Table 18*

demonstrates the keywords are far more frequent in this small specialised corpus than in a large-scale reference corpus.

Table 18: Frequency of nouns, adjectives and full verbs related to the BNC

Key nouns (Corpus vs BNC)		Key adjectives (Corpus vs BNC)		Key main verbs (Corpus vs BNC)	
sightseeing	970.00	panoramic	291.00	explore	110.50
tourism	776.00	trendy	162.00	attract	75.33
tours	637.00	breathtaking	151.00	unveil	65.00
options	421.00	culinary	129.00	aquaints	65.00
website	291.00	cosy	129.00	said	65.00
locals	259.00	frugal	129.00	opined	65.00
travel	259.00	bustling	129.00	teaches	65.00
artwork	226.00	cobbled	97.00	knew	65.00
attractions	204.89	cheapest	97.00	dedicate	65.00
basilica	194.00	affordable	97.00	boasts	64.50

The high recurrence of *website* indicates that city tours are promoted mainly via direct internet marketing, and with adjectives like *frugal*, *cheap* and *affordable* tourists are encouraged to buy them.

(1) *Madrid City Tours is an <u>online travel website</u> featuring Madrid excursions, including trips and tours, along with information on hotels and accommodations.*

(2) *If you visit Moscow choose us for Kremlin tour, we provide the best service for an <u>affordable price.</u>*

Although the frequency of adjectives is considerably higher than that of main verbs, which is another characteristic feature of tourism texts**,** the use of seemingly rare verbs in general English like *unveil*, a*cquaint*, *dedicate* or *boast* make the city tours more attractive, inspiring the participation in sightseeing.

(3) *The French capital <u>boasts </u>the most famous and expensive hotels on the planet.*

(4) *View this cosmopolitan city from the comfort of an air-conditioned bus while a professional tour guide <u>acquaints</u> you with its most popular and historic sites.*

5.7.3. The use of emphatic language

Gotti (2006, pp. 26-29) considers emphatic language use a distinctive feature of tourism discourse compared with other languages for specific purposes. The characteristically predominant persuasive function abounds in the use of nouns, adjectives and expressions that convey hyperbolic and positive images of destinations and services in tourism texts. This concept is in correspondence with euphoria, a tendency of tourism discourse 'to speak only in positive and glowing terms of service and attractions it seeks to promote (Dann, 1996, p. 65) 'through the use of superlatives, hyperboles and different linguistic devices to emphasize the uniqueness of the tourism entity.

(5) *Clusters of towering, 100-year-old palm trees provide pockets of shade from the Florida sun. A breeze wafts softly through the park, a touch of salty coolness from the nearby Atlantic.*

(6) *The scenery was amazing and the guides were professional, well versed in their field and extremely knowledgeable. The lunch stop was great as was the butterfly ranch…"*

(7) *With over 17,000 islands and a collection of glorious beaches, pristine lakes and rushing rivers unmatched anywhere in the world, Indonesia is the perfect place to take a splash and embrace some exciting water sport.*

5.7.4. Means of persuasion

Descriptive and evaluative adjectives

Since all the adjectives in the frequency list (Table 2) have positive connotation, the descriptive adjectives like *panoramic*, *breath-taking* and *bustling* help create a captivating and vivid imagery of cities, while *culinary*, *cosy*, or *cobbled* evoke longing and nostalgic feelings of the atmosphere of old city centres.

(8) *You will discover a fascinating city with Arab origins, spectacular architecture, bustling modern parts, and great nightlife.*

(9) *Your guide can take you through the charming cobbled streets around San Jacinto Plaza to browse its thriving art market.*

As presented in the extracts, descriptive and evaluative adjectives trigger positive and lively images of the destinations and experiences in the minds of the prospective visitors. Cook (1994, p. 11) argues that readers use schemata in discourse processing, that is, mental representations of typical instances and situations, in order to 'predict and make sense of the particular instance which the discourse describes'. The extensive use of descriptive and evaluative adjectives, especially superlatives, communicates a really engaging image of the venues, enhancing their extraordinariness, and increasing the temptation to visit them. The descriptive and evaluative attributes in the following examples convey the message that the attractions are must-see and the experiences are not-to-be-missed.

(10) *The <u>diverse architectural</u> styles and <u>lavish</u> landscaping provide a <u>truly breath-taking</u> sight.*

(11) *We take you on an <u>unforgettable bilingual narrated</u> sightseeing cruise to the <u>scenic</u> bay.*

(12) *A <u>bonafide culinary destination</u> with sophisticated restaurant option for every taste, the city's every gourmand's heaven.*

In the last extract the professional client-oriented attitude, the indulging culinary experience, and the impeccable service are strengthened with an implicitly convincing attribute, *bonafide*.

Intensifying adjectives

Some of the intensifying adjectives of the articles have a general heightening effect on the nouns that they pre-modify like *true* (9), *great* (56) or *real* (6), and some of them function as amplifiers denoting a high degree as they are inherently superlative, such as *perfect* (24), *unique* (26), or *superb* (10). *Table 19* shows the most frequent intensifying adjectives.

Table 19: *Frequency of intensifying adjectives*

Rank	Intensifying adjective	Frequency
1	great	56
2	unique	26
3	perfect	24
4	superb	10
5	true	9
6	real	6
7	absolute	2

Total: 133

The most frequent intensifying attribute, *great*, accentuates the excellence of venues, culinary experiences, and facilities. The second most favoured adjective, *unique,* refers to locations, cultures and attractions, while *superb* emphasises the fineness of cuisines and vistas. The intensifying adjective *absolute* (2) occurs with collocates that also add emphasis to the nouns that they modify; *the absolute best way to*, *the absolute must-sees*.

Sequences of adjectives

In strings of adjectives the multiplicity of attributes reinforces the positive impression of the described tourism product, however, the use of threadbare words makes these descriptions sound like clichés. As compensation, the use of some buzzwords in the sequences makes the attractions more appealing.

(13) *Just wonderful… really beautiful, peaceful and cultural city.*

(14) *Mexico City is really amazing, diverse, beautiful, mind-blowing and thought-provoking.*

As the latter excerpt demonstrates, exaggeration is accepted when these sequences help visualise the attraction and arouse the desire to take part in a once-in-a-life-time experience.

Superlatives

Whereas deficiency is concealed in tourism texts, 'salience', a concept introduced by Kress and van Leeuwen (1996, p. 183), helps to highlight the positive aspects of a tourism product, this way increases the degree of persuasion. Besides the high recurrence of keywords with positive connotation, salience is best achieved with the use of superlative attributes, as these forms express the highest value of the described entities distinguishing it from others, and emphasising their uniqueness.

Examining the occurrence of superlative forms, the total number of superlatives found with wildcard is 187; the '*the most **' form is the most frequent with 88 instances, the irregular '*the best*' has 55 occurrences, and there are 44 '*the *est*' forms. In addition, some amusing serendipities turned up when searching for superlatives with 'est' suffix such as 'west', or 'guest'. The majority of '*the most **' superlatives describe cities, places of interest and tourist attractions. '*The best'* refers mainly to places, catering, hotels, prices, and means of sightseeing. The frequency of occurrence of the superlative forms is demonstrated in *Table 20*.

Table 20: *Frequency of superlative forms*

RANK	SUPERLATIVE	FREQUENCY
1	'the most *'	88
2	'the best'	55
3	'the *est'	44

Total: 187

These superlative adjectives reflect the alluring and superb nature of tourism language; they give additional emphasis to the uniqueness of the described destinations or services.

Comparisons

Comparison as a persuasive communicative technique is employed in tourism language either to emphasize the excellence of the destination or service related to others, or to highlight their exceptional qualities in order to diminish the effects of strangeness in tourists. Similes and metaphors, as the most powerful devices for comparison, depict the tourism product with exaggeration and, in doing so they change people's attitudes. As illustrated in the following examples, similes help tourists become familiar with the destinations as they are likened to something well-known and pleasant.

(15) *This place has a warm and intimate atmosphere like something out of the Arabian Nights. You drive into the middle of nowhere and suddenly this mirage of crenellated walls and towers appears on top of a sand dune.*

(16) *Adorned with huge skyscrapers and pristine beaches, Dubai attracts tourism like honey a group of bumblebees.*

'Magic' is a key metaphor of tourism, implying that a touristic experience is too mystical to be real or true; however, it sparks the imagination of tourists and arouses the desire to visit that 'dreamland'.

(17) *Come and explore some of the most magnificent piazzas, ancient buildings, impressive fountains and incredible views that unveil the magic and the splendour of the eternal city!*

(18) _It's like living in dreamland_! _The interior is out of this world beautiful and so are the outdoors with all the trees and gardens! Magical!_

Besides creating an emotional link to the addressed tourist, a metaphor with an exclamation mark elicits an immediate urge not to miss these enchanting sceneries.

Intensifiers

Intensifiers, as emphatic devices strengthen the meaning of the words that they modify, as well as give them an additional emotional context that enhances their attractiveness. The most frequent intensifier in the articles is *very* with 40 cluster types that highlight the advantages of city breaks and city tours. Other frequent intensifiers are *really* with seven occurrences, and *quite* with six instances.

Epistemic adverbial 'certainly'

The use of the epistemic adverbial 'certainly' indicates that a statement is accepted as valid and evident. Kress and van Leeuwen (1996, p. 160) accentuated that 'the truth value or credibility of linguistically realized statements about the world' is an important aspect of the representation of reality. The adverbial 'certainly' expresses different meanings and fulfils different discourse functions in tourism texts, it is used to emphasize that something is true without doubt, or it is extremely likely to happen. Three main persuasive functions of 'certainly' could be identified in travel articles; expression of opinion, attitude, and emotion.

- **Expression of opinion**

(19) *It is probably the most important building in Barcelona and certainly the most visited.*

- **Expression of attitude**

(20) *The metropolis can never be disappointing as all those taking flights to Rio de Janeiro can certainly expect a lot to explore, be it the frolicking carnivals, the beautiful landscape or the beach culture.*

- **Expression of emotions**

(21) *London is certainly one of the most charming cities that one can visit during their most valuable vacation.*

The discourse functions of 'certainly' illustrated in the examples help to testify that the information is fact-based and these venues are worth visiting.

Imperatives

Imperatives are used to urge someone to do or not to miss something, giving an impetus to tourists to become participants of an exceptional experience.

(22) *Experience the best of 10 European cities on a save-or-splurge budget.*
(23) *Remember to make sure to ride on a bus when visiting the canyon to enjoy the views, shopping centers, as well as the relaxing ambience along the rim trail.*

Exclamations

Contrasting a negative feature, which no means can be concealed, to a captivating image lessens the degree of deficiency with the use of an exclamation mark, as well as heightens the attractiveness of the place.

(24) *Vibrancy juxtaposed against economic squalor! A city of plentiful opportunities, the Mumbai dream!*

The use of exclamation mark not only calls the reader's attention, but also emphasizes the excellence of the tour operator.

(25) If you want to make the most of your Grand Canyon Bus Tours, we can offer the most comfortable trip for you and your family!

Infringement of taboo

In order to attract more people and amplify temptation, instead of conventional descriptions, cliches and stereotypes, the infraction of taboo has an intense attention drawing role:

(26) 'A lot of people just down beer and don't think about what they're downing, like they did in college'

(27) The 'Polygamy Experience' bus tour in Colorado City takes visitors to a community in the Utah-Arizona border where plural marriage is still practised. Travellers are introduced to men in the community who typically take one 'legal' wife and additional 'spiritual' wives. The four-hour tour is run by Richard Holm, who once had three wives.

Person pronouns

Person pronoun 'I'

The accounts of tours and journeys as personal experiences give the impression of authenticity as the use of the personal pronouns 'I' and 'we' increases their truthfulness. In Jorgensen's (2004, p. 73) view the subjective representation of a destination is likely to result in increased number of visitors.

(28) <u>I spent about 4 hours </u>inside this wonderful museum, <u>amazed by the works of art and antiquities.</u>

(29) <u>I recommend it to everyone who wants to relax</u>, eat delicious foods and visit a classical city with unique monuments.

Pronouns 'you' and 'we'

The discourse strategy of ego-targeting is often applied in tourism texts to persuade tourists by emphasizing their uniqueness as personalities (Nigro, 2006, p. 188). Addressing them with '*you*' and distancing '*us*' from others will give them the feeling that they are special and unique as the focus of attention is directed towards them.

(30) *Truly speaking, Taj Mahal is amazingly beautiful and <u>once you see</u> this marvellous monument also referred to as the epitome of love, <u>you are sure to get spell bound</u>.*

(31) *As <u>we have always shared a very trusted rapport</u> with our tour providers, <u>we have managed to offer </u>some of the best deals on hotels that can be booked via this new feature and help the travellers with finding convenient and comfortable accommodation.*

In the above listed instances personal accounts, recommendations, and the strategy of ego-targeting are applied as persuading promotional marketing tools, to amplify the reliability and trustworthiness of the information.

5.8. Conclusion

This corpus-based analysis of the use of persuasive discursive techniques in travel articles shed light on some specific features of tourism discourse, and presented some ways how lexico-grammatical choices as inherent persuasive devices affect customers' perception and attitude. Style, lexis, syntactic structures, and persuasive discourse techniques such as self-presentation

of travel experiences, or individualisation with ego-targeting were detected as powerful means of persuasion. These discursive devices trigger attractive imageries in prospective tourists' minds, and urge them to experience the excitement that they foreshadow. Lexical choices, such as the use of keywords and keying, loads of attributes, comparisons, and metaphors, tend to exaggerate in order to increase their persuasive power, and highlight the positive aspects of touristic experiences. The use of certain syntactic structures in the descriptions of tourism products, such as attributive or intensifying adverbial clusters, despite the exaggerated and hyperbolic images, conveys the message that they are authentic representations of reality.

From a pedagogical perspective, the implication of the findings in teaching the language and discourse of tourism to ESP students can be the target of further research, with special emphasis on the development of language learners' professional discursive competence.

CHAPTER SIX

LEXICAL BUNDLES AND COLLOCATIONS IN THE TOURISM ENGLISH CORPUS

6.1. Introduction

The selection of lexical phrases is one of the greatest challenges of vocabulary teaching, formulaic language makes language teaching and learning more demanding. Since 'lexical bundles are not idiomatic in meaning, and not perceptually salient' (Biber & Barbieri, 2007, p. 269), learners tend to have difficulty in using them. Additionally, as Willis (2003, p. 166) remarked, 'pedagogically the main problem with phrases is that there are so many of them'. It is essential that learners notice and observe lexical chunks in authentic context, recall and retrieve them in language production, therefore, the acquisition of formulaic language has to be an organizing principle of language teaching.

In the Lexical Approach prefabricated items constitute the core of word knowledge, thus enhancing learners' awareness of formulaic language is a way of fostering their language development. As multiword units are stored in and retrieved holistically from the mental lexicon they allow for the production of fluent spoken discourse; 'formulaic language is intrinsically connected with functional, fluent, and communicative language use (Schmitt, 2005, p. 16). Nevertheless, as Wray (2002/b) observed, formulaic sequences seem to contribute to the mastery of grammatical structures only to a certain extent, therefore, she advises that in a sensible course' a balance should be maintained between formulacity and creativity (p. 148)'.

This chapter gives an overview of the corpus-based approach to formulaic language, presents the processes and methods of identifying lexical bundles and collocations, and describes the compilation of collocation lists. It discusses various aspects of formulaic language including its terminology, taxonomy, different approaches to the identification of lexical bundles and collocations, as well as the pros and cons of different statistical measures of collocation analysis. The findings of the structural and functional analyses of multiword units and the

collocational analyses are presented, and their pedagogical applications are delineated and illustrated with a set of tasks.

6.2. Corpus-based approach to formulaic language

6.2.1. Formulaic language

The term *formulaic language* covers a range of multiword units including idioms, collocations, phrasal verbs and other fixed or semi-fixed expressions that facilitate fluent native-like language processing and use (formulaic language and collocations are discussed in 2.4.1. and 2.4.2.). The multiword structures of formulaic language carry a conventionalized 'holistic' meaning that does not derive from the meaning of their components. Formulaic sequences differ along various dimensions, such as their degree of fixedness, their shematicity, that is, whether they allow to fill open 'slots' with variant lexical items.

'In English, there are many multi-word expressions that function as a structural or semantic unit' (Biber et al., 1999, p. 988). On the grounds of their idiomacity and fixedness, different kinds of multi-word units can be distinguished including idioms, collocations, lexical bundles and lexico-grammatical associations. Lexical bundles as recurrent expressions are more common than idioms, these sequences of words generally appear together in natural discourse, making it more coherent. Whereas idioms are relatively invariable expressions, lexical bundles are not fixed units, and, in general, they are not complete structurally. Nevertheless, lexical bundles can be regarded as collocations since they have a tendency to co-occur (Biber et al., 1999), but differ from collocations in the way that they are distinguished mainly on the basis of high-frequency, rather than on their structures or phraseological constraints. As multiword units represent recurring sequences in a collection of texts, with corpus methods and user-friendly corpus tools a frequency-based analysis relatively easily can be carried out.

The diversity of multiword sequences is manifested in the wide range of terminology used for various kinds of formulaic language, including *chunks, collocations, conventionalised forms, clusters, formulaic speech, formulas, holophrases, multiword units, prefabricated routines,* or *ready-made utterances* (Carter, 2012, p. 2). Yet, this list is not complete, other terms referring to formulaic language include *lexical bundles* (Biber et al., 2004)*, n-grams* (Gries, 2008), or

phraseologisms (Gries, 2008). Mainly the term *lexical bundles* is used across this study, interchangeably with *lexical chunks*, *clusters*, *multiword- units* and *sequences*.

Lewis (1997) categorized lexical units as *words* (journal), *polywords* (as a matter of fact), and *collocations/word partnerships* (e.g., basic principle or hold a stance). Lexical bundles can be defined as recurrent formulaic sequences; however, they frequently do not form structural units, and although some of them are considered extended collocations, they are different as they incorporate function words as well. Concerning collocations there is a semantic relation between the node and its collocates, whereas in lexical bundles the syntagmatic relations, colligations are characteristic (Cortes, 2015). Collocations are commonly regarded as subsets of formulaic language (Wray, 2002/b); nevertheless, the *Lexical Approach* considers *lexical chunks* and *collocations* different notions. The term lexical chunk refers to pairs or groups of words that appear in combination, and although collocations are included in this category, they are defined as sets of lexical content words that are commonly observed together. Lexical bundles that are generally associated with collocations have become important subject matter of lexical studies (Biber & Conrad, 1999; Biber, 2003; Stubbs, 2002).

Formulaic sequences can be long or short, they can be used for different purposes such as to express a message or idea, to give information, or fulfil different pragmatic functions *e.g.,* declining an offer. Fixedness is a key feature of formulaic language, allowing to be memorized and used as whole units. However, formulaic language is not completely fixed, in fact allows for flexibility and variations too, as in *burn your boats* or *burn your bridges*. When a piece of formulaic language becomes well-known by the members of a speech community, it can be creatively altered, yet it remains comprehensible, such as *break the ice*, *ice-breaker*, or *ice-breaking* (Carter, 2012, pp. 3-4).

Corpus-based explorations reveal interesting patterns of formulaic language; formulaic sequences can be characteristic to different genres (Weinert, 2011). Specific collocations might be associated with particular genres, registers or texts, as the collocations of special domains routinely combine certain nouns with certain verbs or adjectives. Biber (2003) found that lexical units are reliable indicators of different registers, and the distribution of lexical bundles in various registers are markedly different; in the register of conversation nearly 90 per cent of the lexical bundles are declarative or interrogative clause fragments, whereas in academic prose the bundles are basically phrasal segments.

Surveying research on formulaic language Gries (2008) identified six variables that help define its terminology.

1. *The types of linguistic units involved in the phraseologism*. Generally, research on formulaic language focuses on co-occurrences of words; in contrast, Gries' definition involves the possibility of co-occurrences between one or more words and grammatical structures.

2. *How many units are involved.* Collocation extractions are often based on bigrams (co-occurrences of two elements), since they can be easily detected and examined statistically, however, this perspective is not preferable, as clusters can be of any length.

3. *How frequent the co-occurrence of units must be to consider it as phraseologism.* Lexical association measures indicate the statistical strength of word co-occurrences, many of them are designed to identify not only frequent sequences, but also low-frequency salient ones, like idioms. However, researchers may determine formulaic language merely on the basis of semantic or syntactic criteria, or decide intuitively what is formulaic.

4. *How much syntagmatic distance can be between the units.* The extracted multiword units, generally bigrams, often focus only on adjacent elements; however, there can be one or more internal gaps between the constituents.

5. *The extent of lexical and syntactic flexibility among the units*. A certain degree of lexical and syntactic inflexibility is a characteristic feature of formulaic language. A sequence of words is typically considered formulaic if it sounds more conventional and native-like than an alternative with the same intended meaning. As an instance, in English *strong coffee* is habitually used, while *powerful coffee* is not; although both sequences are compositional, native-like selection delimits the lexical choices. Some multiword units may include partially-fixed paradigmatic slots where words can be replaced.

6. *Whether or not the units exhibit semantic unity and non-composability.* In terms of semantic unity and non-composability there are various approaches to formulaic language. Some researchers regard non-composability a prerequisite for formulaic language, while others categorize formulaic language into different types on the basis of their composability or non-composability, such as idiomatic or non-idiomatic language. In addition, some approaches do not account for semantics at all; frequency-based statistics are the only means of identifying formulaic language.

Gries' taxonomy with its six dimensions provides a principled framework for the synchronization of linguistic approaches to formulaic language.

6.2.2. Approaches to the identification of formulaic expressions

Studies on formulaic language, collocations and lexical co-occurrence have been conducted within the framework of two traditional approaches: the frequency-based tradition represented by Firth (1951) and Sinclair (1991), and the phraseological tradition (Aisenstadt, 1979; Cowie, 1981). Since the late 1990s, owing to the advancement in computer technology, multiword sequences and the discourse functions that they perform have been studied empirically based on the analyses of corpora (Nattinger & DeCarrico, 1992; Hunston & Francis, 2000, Sinclair, Jones & Daley, 2004).

In terms of phraseology collocation refers to a linguistic phenomenon whereby words tend to co-occur, exhibiting idiosyncratic combinatory and semantic properties. Collocations as phraseological units are arbitrary, semantically transparent, but unpredictable in form. The components of collocations represent a different semantic status; the bases or nodes are semantically autonomous, whereas the collocates are determined and selected by their bases. Although collocations undergo semantic specialisation and grammaticalisation to a certain extent, they are different from other types of phraseological units that have a fixed form and non-decomposable unified meaning (Pastor, 2017, p. 29).

In computer-assisted frequency-based approaches of collocations the term has been used to refer to a statistically idiomatic multiword unit or expression, that is, to particular combinations of words that occur with strikingly high frequency. These approaches to collocations consider frequency and co-occurrence distinctive features, regardless their semantically-based or statistically-based nature, however, they are not integrative enough to provide a generally accepted proper definition of collocations (Pastor, 2017, p. 30).

Nevertheless, it should be borne in mind that the findings of collocation extraction are only reflections of what the selected texts of the corpus contain. Therefore, interpretations of corpus evidence should be treated with caution, as it is not possible to attain total linguistic accountability. Castello (2014, p. 7) warns that corpus data helps only infer the underlying meanings of words and decide whether collocations are plausible, thus, 'observations should not be regarded as concrete representations of language use', but as reasonable evidence based on the data available for analysis.

In corpus-based research on collocations generally three criteria have been proposed for identifying collocations; (1) *distance*, (2) *frequency*, and (3) *exclusivity*. The *distance* from the node word specifies the span, also called collocation window that can be as little as one word, *e.g.*, adjectives preceding nouns, or as many as four or five words on both sides of the node when we inquire into more general associations. *Frequency* of use, the second criterion, indicates the typicality of word associations; as an instance *'in time'* is an important chunk in tourism English. As for *exclusivity*, '*in*' occurs with numerous other nouns such as *summer* or *advance*, consequently '*in*' and '*time*' do not have an exclusive relationship. *Time* has a stronger and more exclusive link to *zone*; when *zone* crops up in tourism texts it is highly probable that the preceding word is *time* (Brezina et. al, 2015, p. 140).

In addition to the criteria discussed above, Gries (2013, quoted in Brezina et.al, 2015, p. 140) suggests that three other aspects should be considered among collocates, namely, *directionality*, *dispersion* and *type-token distribution*. *Directionality* indicates that the attraction between the node and collocate is rarely equivalent; the word *zone* has a stronger relationship with *time*, than the word *time* with *zone*, as it frequently co-occurs with other words. *Dispersion* refers to the distribution of the node and the collocate in the corpus, while the *type-token distribution* is a desirable criterion that takes into account the strength of a particular collocational relationship, as well as other collocates around the node word. As collocates of nodes do not occur in isolation, but as parts of a complex network of semantic relations, Brezina *et al.* (2015, pp. 140-141) suggest that a seventh feature, *connectivity*, should be added to these criteria.

Collocation networks, the lexical patterns of a particular field or discourse that collocate with each other, have been investigated in studies on the basis of both general and ESP corpora (Alonso et al., 2011; Williams, 1998). Collocation or lexical networks can give insight into the specific lexical connections of a particular discourse, grounded on multiple comparisons of the interconnectedness of pairs of words. For the automatic identification of collocations and for the exploration of their connections different association measures can be applied such as *Mutual Information* (MI) statistics or the *GraphColl* software.

6.2.3. Collocations and meaning

Cruse (1986, p. 37) challenged the traditional view that idioms are expressions whose meaning cannot be inferred from the meaning of their parts. The conventional view needs to be modified

to apply to collocations; the argument that the collocational patterns define the meaning of the constituent lexical items can be reversed, that is, the particular meaning of an item determines its range in the collocation. Nattinger and DeCarrico (1992, pp. 181-182) share Cruse's view and argue that the collocational range of a lexical item not only contributes to its meaning, but also its collocates 'share meaning to a great extent'. Nevertheless, Lyons (1977, p. 613) warns that to one extreme, it cannot be asserted that the collocations of a lexeme are determined by its meaning/s, or to the other extreme, the meaning of a lexeme cannot be defined only as a set of its collocations.

According to Sinclair (1991, pp. 115-116) the relative frequency of the node and its collocates determines to what extent the collocational relation contributes to the meaning of the node. In his stance only those collocations in which the node occurs with less frequent collocates need 'a semantic analysis of a word'. Nattinger and DeCarrico (1992, p. 20) argue that there is an inverse relation between the predictability of the collocates of a particular node and the meaning that is derived from its collocational relations, that is, fixedness concerning 'mutual expectancy... results in loss of meaning' when the possibility of the 'open-choice' is eliminated.

6.3. Statistical collocation measures

Probably the most popular method for identifying collocations or lexical bundles is the counting of co-occurrences and applying lexical association measures. The most basic statistical measure is a frequency count, and widely-used association measures indicate how much more or less often a particular cluster is observed than might be expected incidentally. Therefore, all collocation measures are based on the comparison of observed frequencies and expected frequencies, yet, they lay different emphasis on different aspects of the collocational relationship (Brezina et al., 2015, p.145).

Church and Hanks (1990) proposed the so-called *node word method* for identifying significant collocations in a corpus with statistical measures. The steps of the process are as follows:

1. for each instance of the node word, find all the collocates occurring in an *n* word span and keep a record of them

2. with statistical methods assess how noteworthy the co-occurring words are

3. sort words based on statistics, include only the items with the highest scores.

The scores of statistical measures are calculated not only by considering the frequency of the target sequences, but by relating the number of occurrences to the size of the corpus. Sometimes the number of occurrences may not convey meaningfully the frequency of a word, thus *normalised* or *relative frequency* indicates how often a word occurs per thousand or million words. Lexical association measures provide more sensitive descriptions of formulaic language than relative frequency, as they can encapsulate sequences that are infrequent yet formulaic.

By means of probability measures, concordances can be complemented with statistical figures that can reveal salient facts about the combinatory preferences of a word. Statistical measures became a standard feature of many corpus-querying tools including the *CLT*, *AntConc*, or *WordSmith Tools;* these corpus software incorporate various kinds of statistically-based collocation listing tools. The probabilities of each component words of a collocation are estimated on the basis of their frequency of occurrence in a corpus (Kilgariff, 2002, pp. 808-809).

Statistics vary in terms of how they assess and measure lexicographic salience (Kilgariff, 2002, p. 808). Statistical measures allow testing the significance of the observed differences between frequency data arising from a corpus; significance tests, such as calculating keywords or collocations assess how likely a particular result is random, or significant that can be relied on. A commonly used significance test in corpus linguistics is the *Chi-squared test*; this non-parametric test compares the differences between the observed frequencies in the data with the expected frequencies, however, Chi-squared test can be unreliable when very small numbers are involved in testing. When searching for the collocates of a word, we test the significance of the co-occurrence of the adjacent or nearby words that appear once or more in a corpus. Generally, the results can be considered significant if there is a 95% chance that they are not brought about by coincidence.

Mutual information (MI) statistic is presumably the most well-known association measure. MI compares the probability of observing two words together with the probability of observing them independently, and tends to select characteristic collocations regardless of the frequency of collocates. However, as MI measures evaluate strength to a greater extent than word association, they often rank very low-frequency high-possibility bigrams too highly, giving unneeded weight to collocates, even when both components of the bigram are hapaxes. The *log-likelihood* statistics is similar to MI, but allows for less reliable estimates of salience based on very low frequency counts, presenting numerous low-frequency collocates.

T-score operates on a similar basis; its extended counts support more accurate estimates of probabilities of co-occurrence. When it is used to measure the salience of one word in relation to another, the standard deviation between the observed and expected frequencies of collocations are measured, providing the independent frequencies of each collocate as well. On the other hand, although T-score lessens the problems associated with MI, it gives high significance to frequently occurring words (Kilgariff, 2002, p. 809).

6.4. Research objectives of the multiword unit and collocation analyses

In order to get a deeper insight into the lexico-grammatical patterns of the TEC texts, I carried out multi-word unit and collocational analyses. Studies on multiword units have been conducted mainly related to academic language (Csomay, 2013; Cortes, 2013; Jablonkai, 2010), research on the lexical bundles of tourism discourse has been published only related to the e-language of hotel websites (Fuster-Márquez, 2014), and tourism collocations have been in the focus of a few studies (Kiss & Horváth, 2016; Pierini, 2009). Therefore, the analyses of multi-word sequences in the TEC texts aimed to identify characteristic bundle types in tourism language.

The aim of collocation analyses was twofold; they were conducted to detect recurring and unfamiliar lexical patterns of tourism language, and to reveal different uses and meanings of certain lexical items. An additional aim was the compilation of lists of tourism-specific three- and four-word bundles, as well as collocation lists of tourism keywords that can serve as starting points for syllabus and course design, and for material development.

Durrant (2008, p. 252) maintained that collocation lists are viable, and they could provide basis for more comprehensive phraseological listings. The reason behind the compilation of collocation lists is their potential application in language teaching; these lists can help enhance language learners' word-meaning and collocational awareness, increase the size, breadth and accuracy of their lexicon, and this way promote the development of fluency. In addition to language learners' vocabulary expansion, due to the increased input provided by collocation lists while learning collocations all four skills are involved; reading and listening activities promote the noticing of collocations, whereas speaking and writing activities create opportunities for practice (Károly, 2005, p. 63). According to Carter and McCarthy (2014) by memorizing collocation groups language learners become aware of certain lexical and

syntagmatic restrictions; 'collocations teach students expectations about which sorts of language can follow from what has preceded (p.75)'.

6.5. Identifying lexical bundles and collocations in tourism texts

6.5.1. Research questions

In order to reveal what sorts of lexico-grammatical structures are used in the texts of the TEC, the present study addresses the following research questions:

1. *What frequent three- and four-word lexical bundles can be identified in the TEC texts?*

2. *How can the identified lexical bundles be described structurally and functionally?*

3. *What are the frequent collocates of the lexical items of the TEC word lists?*

4. *What are the grammatical and semantic preferences of the selected key node words?*

6.5.2. Methodology of the analysis of lexical bundles

A corpus-driven approach was followed in the multi-word unit analyses, and evidently, *lexical bundle* was the unit of analysis. For the detection and extraction of multi-word sequences frequency was the basic criterion, to identify collocational patterns I followed the *node word method* advocated by Church and Hanks (1990), that is, the identification of the word associates of node words were based on statistical measures.

In the first section of the study, I lay emphasis on three- and four-word bundles as three-word clusters are frequently parts of four-word units, in addition, they are more frequent than five-word bundles, and, the two-word chunks were examined as collocations. To identify the most frequent bundles I searched for possibly interesting four-word lexical units; it seemed manageable to classify such bundles in context as they are likely to occur with lower numbers. To reveal the structural and functional characteristics of lexical bundles quantitative and qualitative analyses were carried out.

To search potential three- and four-word bundles the TEC texts were converted to plain text format, and the concordance tools of *AntConc 3.5.8* were used, mainly the Clusters/N-gram function. On the basis of the output of the corpus software, lists of three- and four-word lexical

bundles were compiled, each lexical bundle of both lists was manually checked and selected to include.

The procedures of the analyses of lexical bundles are as follows:

1. The multi-word units were identified applying the Clusters/N-gram function of *AntConc 3.5.8*.

2. Lexical bundles were selected based on their frequency and subject-specificity; frequency lists of three- and four-word bundles and a list of frequent tourism-specific clusters were compiled.

3. Qualitative analyses of the multiword units were carried out; the structural and functional features of three-and four-word bundles were investigated.

6.5.3. Identifying lexical bundles

6.5.3.1. Structural types of lexical bundles

Lexical bundles including collocations are identified mainly on the basis of *frequency*, as highlighted in 6.2.2. and 6.3., and the identified lexical units are commonly classified on the grounds of their grammatical structure. I employed Biber and his fellow researchers' (1999; 2004) structural taxonomy to determine lexical bundles in tourism texts. The three main structural categories in Biber *et al*.'s classification involve (a) lexical bundles that incorporate *noun phrase-based* (NP) or *prepositional phrase-based* (PP) fragments, (b) lexical bundles that incorporate verb phrase fragments (VP), and (c) lexical bundles that incorporate dependent clause fragments, such as *which is the* or *if you look at*.

Each of the main bundle types involve several subcategories as illustrated in *Table 21*. NP-based clusters include noun phrases with *of-phrase*, e.g., *the rate of the*, or post-modifier fragment as *the room in which*. PP-based bundles consist of prepositional phrases with of-phrase, for example *on the top of*, or other fragments like *at your service* or *on the menu*. VP-based bundles are multiword sequences with a verb component, such as *it is cancelled* or *it will be refund*. The sub-type of lexical bundles incorporating different clause fragments, may include *1st/2nd person pronoun* and *dependent clause* fragments, *WH-clause* fragments, *if-clause*, *to-clause* or *that-clause* fragments.

Table 21: Structural types of lexical bundles (Biber et al ., 2004, p. 381)

Structural types	Sub-types	Sample bundles
1. Lexical bundles that incorporate verb phrase fragments	1.a 1st/2nd person pronoun +VP fragment	*I'm not going to*
	1.b 3rd person pronoun + VP fragment	*and this is a*
	1.c discourse marker + VP fragment	*I mean I don't*
	1.d Verb-phrase (with non-passive verb)	*have a lot of*
	1.e Verb-phrase (with passive verb)	*is based on the*
	1.f yes-no question fragments	*are you going to*
	1.g WH-question fragments	*what do you think*
2. Lexical bundles that incorporate dependent clause fragments	2. a 1st/2nd person pronoun + dependent clause fragment	*I want you to*
	2.b WH-clause fragments	*when we get to*
	2.c. *if*-clause fragment	*if we look at*
	2.d *to*-clause fragment	*to be able to*
	2.e *that*- clause fragment	*that this is a*
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments	3.a Noun-phrase with *of*-phrase fragment	*one of the things*
	3.b Noun-phrase with other post-modifier fragment	*the way in which*
	3.c Other noun-phrase expressions	*a little bit more*
	3.d Prepositional phrase expressions	*at the end of*
	3.e Comparative expressions	*as well as the*

Besides these structural categories, during the analysis two other types cropped up; bundles that incorporate adjectives and adverbs, and bundles including numbers. The numbers in these clusters are marked with a hashtag # symbol.

6.5.3.2. An inventory of three- and four-word lexical bundles in the Tourism English Corpus

Applying the Clusters/N-Grams function of *AntConc 3.5.8*, a total of 16,579 types of four-word bundles, and 14,579 three-word bundle types were generated. As shown in *Table 22*, examining the top twenty most frequent three-and four-word span bundles, there is some overlap between them as several four-word bundles involve three-word bundles, these multiword units are italicized.

Table 22: Lists of the first twenty most frequent three- and four-word bundles

RANK	Three-word	Frequency	Four-word	Frequency
1	*one of the*	267	*is one of the*	73
2	*some of the*	146	*one of the most*	61
3	*if you are*	113	*one of the best*	58
4	*of the most*	108	*some of the most*	41
5	*as well as*	95	*in the heart of*	35
6	*of the best*	86	*if you want to*	34
7	*is one of*	82	all over the world	31
8	of the city	72	from all over the	26
9	in the world	62	*some of the best*	21
10	it is a	61	*as one of the*	20
11	the food was	57	you are looking for	19
12	a lot of	50	a great place to	18
13	*the heart of*	49	is located in the	18
14	*you want to*	49	*are some of the*	17
15	the food is	47	at the end of	17
16	the city is	44	*if you are looking*	17
17	there is a	44	*of the most popular*	16
18	this is a	44	*the heart of the*	16
19	view of the	44	will definitely be back	16
20	and it was	43	*as well as a*	15

The majority of these bundles are not complete units but rather fragments, often with embedded fragments, however, they have fairly strong grammatical correlations. Concerning their internal relations and constituency there is no idiomatic three-word unit among the top forty most frequent chunks, as their meaning is conveyed by the meaning of their constituents, and there is only one four-word bundle that can be considered idiomatic, *in the heart of*.

The most frequent three-word fragment *'one of the'* (267) is embedded in the most frequent four-word bundle *'is one of the'* (73), and there are three other four-word chunks that comprise this three-word segment; *one of the most* (61), *one of the best* (58), and *as one of the* (20). Their frequency of occurrence is marked in parentheses. Nine three-word chunks from the top twenty frequency list occur in 13 four-word units; the most frequent *'one of the'* is incorporated into 4 four-word bundles, and the second cluster, *some of the*, appears twice in the four-word list.

6.5.3.3. Structural categories of lexical bundles in the Tourism English Corpus

Table 23 presents the structural types of four-word bundles identified in the TEC texts, based on Biber and colleagues' (2004) taxonomy. The structural categories of lexical bundles were examined manually, then corresponding lexical units were rendered to them.

Table 23: *Structural types and sample lexical bundles in the TEC texts*

Structural types	Sub-types	Sample bundles
1. Lexical bundles that incorporate verb phrase fragments	1.a 1st/2nd person pronoun +VP fragment	*I'm not going to*
	1.b 3rd person pronoun + VP fragment	*and this is a*
	1.c discourse marker + VP fragment	*I mean I don't*
	1.d Verb-phrase (with non-passive verb)	*have a lot of*
	1.e Verb-phrase (with passive verb)	*is based on the*
	1.f yes-no question fragments	*are you going to*
	1.g WH-question fragments	*what do you think*
2. Lexical bundles that incorporate dependent clause fragments	2.a 1st/2nd person pronoun + dependent clause fragment	*I want you to*
	2.b WH-clause fragments	*when we get to*
	2.c. *if*-clause fragment	*if we look at*
	2.d *to*-clause fragment	*to be able to*
	2.e *that*- clause fragment	*that this is a*
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments	3.a Noun-phrase with *of*-phrase fragment	*one of the things*
	3.b Noun-phrase with other post-modifier fragment	*the way in which*
	3.c Other noun-phrase expressions	*a little bit more*
	3.d Prepositional phrase expressions	*at the end of*
	3.e Comparative expressions	*as well as the*

Comparing the structural types of the TEC with those of listed in Biber et al.'s classification, the persuasive nature of tourism texts was the reason of adjoining two bundle categories; bundles that incorporate adjectives/adverbs and numbers. In the former category, owing to the descriptive and evaluative features of tourism language, the adjectival sub-type is more frequent than the adverbial. Lexical units that include numerals refer to times, rates and groups of people in tourism texts.

The most frequent structural types are lexical bundles incorporating VP and NP/ PP fragments, indicating that the bundles of the investigated tourism registers, namely, travel articles, tourism adverts and customer reviews, include more phrasal than clausal segments. The broadest structural type category, multiword units that incorporate VP fragments, involves the most sub-

types; 'person pronoun + VP fragment' (*I've ever had*), 'copula *be* + noun/adjective phrase' (*is one of the*), and 'anticipatory *it* + VP/adjective P' (*it is advisable to)*, are frequently occurring ones in the TEC texts. According to Biber et al. (1999) the 'anticipatory *it* + VP/adjective P' fragment that initiates extraposed structures falls into two subcategories; bundles that are controlled by either an adjectival phrase, or by a verb phrase. Examining the *it* + VP/adjective fragments, it turned out that the majority of the word combinations in the TEC are adjectival, and a low number of phrases have a verb complement.

In the category of lexical bundles that incorporate NP fragments two sub-types are prevalent in the TEC texts; noun-phrase with *of*-phrase fragment, and noun phrase with another post-modifier fragment (a). A noun-phrase with *of*-phrase fragment is used to convey a broad range of meanings including physical description and identification of places or quantities. In a noun-phrase with *of*-phrase pattern, '*of*' is sometimes followed by a *Wh*-clause (b).

(a) **The** *rooms have* **luxurious facilities, each equipped with** *a 32-inch flat-screen TV.*

(b) *Parents have to accompany their child to eight particular statues the* **names of which** *we will provide.*

A PP generally occurs as pre- or post-modifier of a noun-phrase, or it clarifies the reference of a nearby pronoun. PP phrases are common in post-modifier structures in the TEC (c), in some cases prepositional expressions are made up of sequences of prepositional phrases (d).

(c) *Unfortunately, we can only imagine how beautiful* **the garden behind the house** *can be in summer.*

(d) *This resort is located* **in the southernmost part of Bali with its own private section of the beach.**

Scrutinizing bundles with dependent clause fragments, *to*-clause and *WH*-clause fragments can be detected most frequently. A great number of *to*-clause fragments start with *'to explore'* and *'to offer'*, two instances, (e) and (f), illustrate their structural relations.

(e) *Holidays in Goa also provide tourists wonderful opportunity* **to explore** *the rich flora and fauna, and historical monuments of the state.*

(f) *Earlier, Goa was prominent for its natural exquisiteness and breathtaking beaches, intended* **to offer** *a rejuvenating experience to the holidaymakers.*

That-clause fragments, lexical bundles that follow a main clause verb, are infrequent in the TEC, the word '*that*' occurs most of the times as a demonstrative.

6.5.4. Functional classification of lexical bundles

6.5.4.1. Functional taxonomies

Biber and fellow researchers (2004, p. 381) distinguished three major categories of multiword sequences in their functional taxonomy, based on the analysis of lexical bundles in academic context, in classroom teaching, and course books;

- <u>stance expressions</u> that convey someone's judgement on certainty, proposition, or ability (e.g., *it is possible to*, *it is necessary to*, *it is difficult to*)
- <u>discourse organizers</u> that are used to structure texts (e.g., *last but not least*, *we can see that*)
- <u>referential expressions,</u> such as attributes, time/place reference (e.g., *one of the most, at the same time*)

Hyland (2008) modified Biber et al.'s (2004) categorisation and introduced sub-categories that reflected on academic genres and research writing, and distinguished them as to whether they function as research-oriented, text- or participant-oriented sequences. Cortes (2002) differentiated four functional categories in her taxonomy; *location markers* (e.g., the other side of), *temporal markers* (e.g., at the same time), *text markers to guide the reader to certain parts of a text* (e.g., the rest of the) and *special use bundles* (e.g., in the form of). Her (2013) data-driven investigations on lexical bundles in research articles reaffirmed Biber's (2003) and Hyland's (2008) claim that lexical units are register-bound.

6.5.4.2. Functional categories of lexical bundles in the Tourism English Corpus

The functional analysis of multiword units was carried out on the basis of Biber et al.'s (2004) and Cortes' (2002) taxonomy. Concerning the analysis of discourse functions, the TEC provides only subtle amount of data for identifying stance-bundles or discourse organizers, and, on the other hand, tourism texts are generally structured only to a certain extent, thus they do not tend to use text organizers, expressions to introduce or elaborate a topic. However, referential bundles including location markers, temporal markers, or quantifying expressions such as *a few of the* or *the plenty of* (Biber et al., 2004; Cortes, 2002) are widely used in tourism texts. An additional category, subject-specific bundles, such as *the driving tours of* or *will definitely be back,* can be detected in large numbers in the corpus. The discourse function categories of bundles identified in the TEC are shown in *Table 24*.

Table 24: *Types of discourse functions and sample lexical bundles in the TEC*

CATEGORIES	SUB-CATEGORIES	SAMPLE BUNDLES IN THE TEC

I. Stance bundles	**A. Epistemic stance**
	
	B. Attitudinal/Modality
	
	  B.1. Obligation/directive
	  B.2. Intention/prediction
	  B.3. Ability
		the fact that the, I don't know
		
		you don't have to, if you want to
		you will enjoy the
		to be able to

II. Discourse organizers	**A. Topic introduction** **B. Topic elaboration/clarification**	*I would like to begin, I want to share my impression* *as well as the, on the other hand, I want to mention*

III. Referential bundles	**A. Identification/focus**
	B. Specification of attributes
	  B.1. Quantity specification
	  B.2. Quality specification
	C. Time/place/text reference
	  C.1. Time reference
	  C.2. Place reference
	  C.3. Text reference
		is considered as the
		
		a little bit of, a great deal of
		same high quality of
		
		opened in July #
		in the Big Apple
		as shown in the

IV. Subject-specific bundles	**A. Reference to participants of tourism** **B. Reference to tourism products**	*overnight visitors from mainland, travellers that need real* *each dish was unique, the average room size*

As it can be seen in *Table 24* altogether 14 functional bundle types could be identified in the TEC texts. *Stance bundles* function as expressions of attitude or certainty, thus, they are used to express *Epistemic stance* and *Attitudinal stance.* Both personal and impersonal categories of *Stance bundles* occur in the TEC, that is, bundles that either involve person pronouns or not. *Epistemic* stance bundles comment on the degree of certainty as *the fact that the, I don't know*, there are only a few instances of this type of bundle in the corpus. *Attitudinal/modality stance* bundles express attitudes towards actions, the three subcategories that could be identified in the corpus express (1) obligation/directive (*you don't have to, if you want to*), (2) intention/prediction (*you will enjoy the*), and (3) ability (*to be able to)*. Concerning the structures of *Stance bundles,* they are mostly expressed with VP fragments.

Discourse organizing bundles allow for introducing new topics and elaborating on them, thus, they fulfil two major discourse functions; *topic introduction/focus* and *topic elaboration/clarification*. Topic introduction bundles either announce a new topic (I *would like to begin, I want to share my impression*) in the tourist accounts, or express intention and desire (*if you would love to, if you look for*) in the TEC texts. Topic elaboration/clarification bundles provide additional explanation and details, such as *I want to mention*, or (*you will*) *understand what I mean*. Some of these bundles are used for comparing or contrasting ideas: *as well as the, on the other hand*. Relating the structural and functional categories, the most frequent structural type of discourse organizer bundles comprise verb phrases, and there are a few bundles that involve dependent clause fragments, adjectives and adverbs, or prepositional phrases.

Referential bundles accounted for the majority of functional bundle types in the TEC. Referential bundles indicate direct reference to the textual context, identify entities or specific attributes that precede a head noun. Three types of referential bundles could be detected in the texts;

(1) identification/focus (*is considered as the*)

(2) specification of attributes including quantity specification (*a little bit of, a great deal of*) and quality specification (*same high quality of*)

(3) time (*opened in July #)*, place (*in the Big Apple)*, or text reference (*as shown in the*).

Concerning the referential functions of the identified bundles, time/place/text reference and quantity specification categories were the most common sub-types. In terms of the interaction of functional and structural categories, referential bundles are mostly PP and NP fragments, some of them include VP and adjectival fragments, and only a few incorporate dependent clause fragments.

Subject-specific tourism bundles constitute a new functional category, comprising multiword units that either refer to participants of tourism (*overnight visitors from mainland, tailor-made for travellers who*), or to tourism products (*each dish was unique, the average room size*). As these bundles refer to and describe tourism entities, these expressions often comprise descriptive and evaluative attributes. The top 20 most frequent tourism-specific four-word bundles can be seen in *Table 25*.

Table 25: *The top 20 most frequent tourism-specific four-word bundles in the Tourism English Corpus*

RANK	TOURISM-SPECIFIC BUNDLES	RANK	TOURISM-SPECIFIC BUNDLES
1	beach resort located on	11	with a view of
2	the food was delicious	12	guided tour in Madrid
3	views of the city	13	a perfect place for
4	you can enjoy the	14	inn suites in Miami
5	indoor and outdoor pools	15	never fails to lure
6	one of the best beaches	16	a delightful and memorable
7	a jet boat ride	17	and explore some of
8	and the service was	18	city tours offer a
9	day tours of Rome	19	destinations in the world
10	tourists from all over	20	drinks and snacks for

Examining the structures of the first 50 most frequent tourism-specific bundles that are displayed in *Appendix K*, it turned out that the majority of them incorporate NP and PP fragments (25), the proportion of bundles incorporating VP fragments are relatively high (16), and, in contrast to the expectations, there are nine bundles that comprise adjectives (8) and adverbs (1). The investigations revealed that tourism-specific bundles are similar in their structure to referential bundles, as both functional categories operate mostly with NP and PP fragments, include VP and adjectival fragments, but do not frequently incorporate adverbial or dependent clause fragments.

6.5.5. Summary of the analyses of lexical bundles

This section gave a brief survey of formulaic language, statistical measures of corpus-based multiword unit and collocation analysis, and the taxonomy of structural types and functional categories of lexical bundles. It presented the findings of the analyses of four-word bundles in the TEC texts in terms of their structural and functional peculiarities. Lexical bundles are basic elements of tourism discourse, the sequences and combinations of words in multiword expressions reflect the specificities of the tourism genre. In regard to the structure of lexical units in the TEC, NP/PP and VP fragments occur most frequently, and the relatively high proportion of bundles incorporating adjectives demonstrates the descriptive and evaluative nature of tourism language. Although VP fragments are characteristic to spoken registers (Biber, 2003), they frequently occur in the TEC texts; convincing narrations of activities,

services, and experiences strengthen the persuasive function of tourism texts. From the aspect of the discourse function categories of lexical bundles, time/place/text referential bundles and quantity specification bundles are the most frequently occurring sub-types, nevertheless, lexical bundles are not often used to express stance, or function as discourse organizers.

6.6. Collocations in the Tourism English Corpus

6.6.1. Methodology of the identification of collocations

Linguistic frameworks such as lexicogrammar, construction grammar and pattern grammar deny the sharp distinction between lexis and grammar, and do not have an exclusive focus on lexical collocations. According to Halliday (1991) lexis and grammar are not end-points of a spectrum, but form a continuum, whereas construction grammar considers language an idiomatic continuum in which constituents can be regarded as pairings of lexico-grammatical forms, as well as semantic and discourse functions. Pattern grammarians Hunston and Francis (2000) describe lexical items in their syntactic environment, asserting that grammatical patterns are often associated with certain lexical instantiations, and lexical items often prefer particular grammatical forms.

Adopting the aforementioned views, the identification of collocations was based on word forms rather than lemmas, as lemmatization may conceal important disparities in the collocational preferences of the different forms of a lemma. To detect collocational patterns, I followed the corpus-driven KWIC method, as well as Sinclair's (1991) practice of observing co-occurrence in a no longer than four-word span. The log-likelihood based keyword method was used to identify frequently occurring collocations, as well as MI scores were employed to select the key collocates of node words.

6.6.2. Procedures of the collocation analyses

To get a deeper insight into the lexical composition of the TEC texts, in the second part of the study tourism collocations were scrutinized. First, frequency lists of key node words were generated, afterwards, the key node words were sorted based on their frequency in alphabetical order. Then, with the Clusters/N-gram functionality of *AntConc 3.5.8* two- and four-word congrams, that is, sets of co-occurring word sequences were generated. As a next step, collocations were identified through the clusters of the concordances, complemented with manual analysis of concordance lines. The concordances were carefully examined to detect

collocations, no minimum frequency was set in order to identify important lexical items with lower frequency. Finally, listings of the most typical binary and four-word span left and right collocates of the items of the TEC wordlist were generated, then the selected key nouns, verbs and adjectives were investigated in terms of their collocational preferences, and grammatical and semantic relations.

Since the TEC corpus could be automatically part-of-speech tagged by the *Sketch Engine* software (Kilgarriff & Tugwell, 2001), search expressions could also be investigated in terms of their grammatical categories and collocations. However, the search patterns of collocations may not include bundles with a number of intervening lexical items, or may not capture multiword units that comprise several function words. Consequently, as the collocation analysis is based on n-gram frequency, mainly continuous co-occurrences could be identified.

To get a more comprehensive insight into the lexical combinations and lexical behaviour of selected items in tourism-specific texts, the investigations focused on providing collocational information for the elements of the TEC wordlist and the selected key node words, as well as on revealing their grammatical and semantic relations. Since semantic preferences are multifarious, and, as there is no established semantic categorization of such preferences of lexical items in collocational research, the semantic prosody and semantic preferences of key node words were investigated on the basis of their semantic relations. The investigated relations include (1) synonymy, (2) antonymy, (3) hyponymy that involves words whose meanings are specific instances of a more general word as *hotel* and *accommodation*, as well as (4) meronymy that denotes a part of its referent, e.g., *main dishes* are meronyms of a *menu*.

6.6.3. Preliminary investigations on attribute-noun type collocations

Prior to the investigations on the possibilities of the compilation of collocation lists and on the grammatical and semantic preferences of key words selected from the TEC, I carried out a corpus-linguistically oriented CLIL study with co-author Horváth (Kiss & Horváth, 2015), in which we investigated the possibilities of the compilation of a high-frequency, pedagogically applicable attribute-noun tourism collocation list, based on the texts of the *Beach holidays* subcorpus. Our study looked at how a corpus can assist students and teachers in the acquisition of collocational competence in CLIL context, and, as both collocations and colligations are needed in ESP and CLIL contexts to achieve collocation competence, the adjective-noun clusters of two keywords, *beach* and *island*, were scrutinized.

6.6.3.1. Data analysis

Following the vocabulary profile analysis, we selected the keywords of the texts in the *Beach holidays* sub-corpus applying AntConc 3.2.4w Keyword function's Chi-square test, then compared the top 100 keywords to statistical Log-likelihood measure findings to increase reliability, and selected the common, shared content words of the two lists. Afterwards we examined the left two-word span adjacent attribute-noun collocates of two keywords with the Clusters function, relying on the mutual information (MI) scores, as they yielded the same results as T-scores. Finally, we compiled a collocation list of the most typical evaluative and descriptive adjective-noun clusters on the basis of frequency and statistical measures.

6.6.3.2. Results

Analysing the vocabulary profile of the corpus (Table 26) a characteristic feature of the specialised tourism texts could be observed; the number of tokens per type is relatively high, accordingly, the type-token ratio is quite low.

Table 26: *The Vocabulary Profile of the Beach holidays subcorpus*

Words in text (tokens):	51150
Different words (types):	4757
Type-token ratio:	0.09
Tokens per type:	10.75

The tokens per type figure shows something very important in this kind of marketing lingo: repetitions. This fortunately coincides with a lexical skill development principle, making the application of this corpus for collocation competence development purposes eminently relevant as well as authentic.

The keyword lists shown in *Table 27* were generated by sorting words on frequency and keyness, the first keyword generating method was the Chi-squared test followed by Log-likelihood ratio analysis. When applying the Chi-squared test keyness is marked with frequency and the keyness and frequency scores are equal, whereas Log-likelihood ratio yields different frequency and keyness scores. However, both Chi-squared and Log-likelihood ratio tests resulted in the same list of keywords as frequency is constant. Only eight content keywords

were found in the top 100 list, as the proportion of function words and proper nouns was extremely high.

Table 27: *Keyword lists of the beach holiday corpus sorted by keyness and frequency (FREQ)*

	Chi-square test				Log-likelihood ratio		
RANK	FREQ	KEYNESS	KEYWORD	RANK	FREQ	KEYNESS	KEYWORD
1	214	214	beach	1	214	296.667	beach
2	46	46	island	2	46	66.542	island
3	43	43	town	3	43	59.611	town
4	42	42	bay	4	42	58.224	bay
5	38	38	ocean	5	38	52.679	ocean
6	34	34	coast	6	34	47.134	coast
7	31	31	resort	7	31	42.975	resort
8	19	19	sea	8	19	26.340	sea

From among the keywords we tested the first two key nouns, beach and island by examining their two-word span left collocates, then selected all the attributive clusters of evaluative and descriptive adjectives, the selective collocation measure was the MI score. The total number of clusters types was 185 related to beach and 84 to island; the number of attributive clusters was 39 pertaining to the former and 24 to the latter. The fifteen most frequent evaluative attributive clusters of these keywords are listed in the order of frequency in *Table 28*.

Table 28: *The fifteen most frequent attributive clusters of the keywords 'beach' and 'island' generated by AntConc3.2.4w Clusters function*

famous beach	**award-winning** island
the best beach	**beautiful** island
the finest beach	**romantic** island
friendly beach	**exotic** island
good beach	*famous* island
luxurious beach	**must-see** island
perfect beach	**unspoilt** island
popular beach	*perfect* island
paradise beach	**extraordinary** island
stunning beach	*friendly* island
ultimate beach	**magical** island
excellent beach	**mystical** island
fabulous beach	*paradise* island
indulging beach	**unique** island
spectacular beach	**wonderful** island

In terms of semantic prosody, the attributes of both keywords evoke positive associations, some of the adjectives like *famous*, *friendly* and *perfect*, printed in italics, occur as attributes of both nouns. The intensifying adjectives, *the best, ultimate, unique* and *extraordinary* amplify the persuasive impact of the described tourism entity. Positive evaluative adjectives such as *excellent*, *fabulous* and *indulging* manifest the well-crafted, captivating feature of tourism texts. Some of them like *famous, friendly* and *perfect*, printed in italics, occur as attributes of both keywords.

Some adjectives can be considered either evaluative or descriptive as they partly denote an objective property, and partly refer to a more subjective stance as *exotic* or *spectacular*. The most frequent attributes refer to the quality, size and function of beaches and islands: *sandy, accessible, long* and *recreational* beaches, and *quiet, large, tropical* islands. It is worth having a look at the least frequent clusters; some collocates of 'beach' like *indulging, oceanfront* or *rated*, and clusters of 'island' such as *unspoilt* or *mystical* rarely occur in tourism texts and course books.

6.6.3.3. The pedagogical implications of attribute-noun type collocations

By observing and analysing authentic texts in a corpus, students can investigate significant or problematic word patterns in natural context, and describe lexical and grammatical features of a language. Concordances are the most visually effective as well as exciting ways of exploring the collocations and colligations in a corpus. *Figure 4* shows some random concordance lines in which the collocates of the keyword *beach* can be observed.

```
  spent relaxing on a quiet beach with soothing, crystal waters
 tely to have to share your beach towel! Read on for a pick of
   of them!    This stunning beach can be found just a short dri
 ecially at sunset.    This beach is without a doubt the option
 n the equally petite Relax beach in Pomos. The beach is usuall
  Relax beach in Pomos. The beach is usually populated with a f
 s a beautiful and secluded beach that offers affordable rental
 s offers affordable rental beach huts (at 500 PhP) for those w
 y or two frolicking by the beach, there are also several resor
```

Figure 6: *Random concordance lines of the keyword 'beach'*

Learners can take a look at some of the attributive collocates of this target word such as *petite*, *stunning*, *secluded,* at the proportion of proper nouns, and a colligational pattern of a sequence of adjectives (affordable rental beach hut). As the observation of concordance lines can be daunting after a while due to the high recurrence of proper nouns and function words, a collocation list provides more targeted exposure to the attributive clusters of the keywords.

Teachers and learners can select frequent attributes to be learnt from the collocation lists as *unspoilt*, *Caribbean* or *paradise* related to the keyword 'island'. The exploration of collocations is not only beneficial in teaching synonymous adjectives like *stunning*, *spectacular* or *fabulous*, special word patterns and the specific lexis and features of the language, but makes the lessons more enjoyable with fun that we, teachers, tend to forget about; serendipities of collocations can serve as a source of humour in ESP and CLIL classes.

The lexical choices and the full-of-promise attributive clusters with markedly descriptive and evaluative adjectives in the Beach holidays subcorpus explicitly refer to the beguiling and sales-oriented nature of tourism language, and outline the register of tourism adverts and customer reviews on sea-sand-sun holidays.

6.6.4. The compilation of the TEC collocation lists and their key findings

6.6.4.1. The TEC collocation lists

Although frequent lexical bundles could be relatively easily identified in the TEC, while focussing on lexical bundles it is likely that important tourism collocations will be overlooked. Creating a listing of two-word tourism collocations and investigations of four-word span sequences that may include positionally-variable expressions could provide an account of high-frequency and genre-specific collocations, as well as basis for their pedagogical descriptions. Applying log-likelihood measure, a ranked list of key collocates was produced, then a number of collocations were manually selected from the concordances, and removed from the list if they included proper names, abbreviations, or acronyms. The most frequent left and right collocates of the items included in the full TEC word list are presented in Appendices L, M, and N.

Having a look at the concordances of tourism collocations it became obvious that a proportion of these units are 'grammatical' collocations, that is, they comprise at least one non-lexical word such as prepositions, determiners, conjunctions, pronouns, and numerals. A listing including high-frequency grammatical collocations has its benefits to learners; as these grammatical collocations lack salience, the most typical word patterns that students need are brought to their attention; thus, they can observe that lexical items tend to 'favour' certain grammatical forms. Concerning lexical collocations, the listing is equivocal; as the views that deny the absolute distinction between lexis and grammar have been adopted, the collocations can be regarded as lexical units that have a grammatical structure.

6.6.4.2. Grammatical behaviour of selected key words

According to the output of the Wordlist function of *Sketch Engine* the majority of the words in the TEC are *nouns* (1,799 items with 47,350 total frequency), *adjectives* (584 items with 17,899 total frequency), and *verbs* (508 items with 25, 989 total frequency). The proportion of the lexical items based on their POS with their total frequency of occurrence is shown in *Table 29* in decreasing order.

Table 29: *The lexical items of the TEC based on their POS with their total frequency of occurrence*

PART-OF-SPEECH	NUMBER OF ITEMS	TOTAL FREQUENCY
noun	1,799	47,350
adjective	584	17,899
verb	508	25, 989
preposition	69	21,776
pronoun	24	10,003
numeral	23	2,142
conjunctions	6	8,578

In order to shed light on the grammatical behaviour of the lexemes in the TEC texts 12 key node words were selected, and the scope and variety of their grammatical relations were investigated. The selected key words include three nouns, verbs and adjectives, respectively:

- Nouns: AIRLINE, PASSENGER, RESERVATION
- Verbs: BOAST, INDULGE, VISIT
- Adjectives: AFFORDABLE, AUTHENTIC, TRADITIONAL

To reveal some characteristics of the grammatical behaviour of the selected key words, the number and variety of their grammatical structure types were investigated with *Sketch Engine*. First, the number and type of the grammatical relations of the selected words were compared in the *TEC* and in the *BNC* as a reference corpus. Although the frequency of words in the *TEC* is not comparable with the normalized frequency of occurrence in the *BNC,* differences in the grammatical behaviour of the target items still can be uncovered in terms of a small specialized corpus and a larger reference corpus.

The number of the grammatical relations of the target words and their identical relations in both *TEC* and *BNC* are illustrated in the order of the frequency of relations in *Table 30*. As can be seen, only three words have the same number of grammatical relations in both corpora, namely, AIRLINE, PASSENGER, and INDULGE, they are marked with shading. As expected, the majority of words exhibit a greater variety of grammatical relations in the BNC than in the TEC, BOAST is the only word that has the same number of identical relations in both corpora, it is highlighted in bold.

Table 30: Number of grammatical relations of selected key words in the TEC and the BNC

Word	Frequency in TEC	Number of grammatical relations in TEC	Number of grammatical relations in BNC	Number of identical grammatical relations
AIRLINE	172	10	10	9
PASSENGER	69	9	9	7
RESERVATION	32	9	6	5
INDULGE	23	8	8	7
VISIT	332	7	13	6
BOAST	**27**	**6**	**9**	**6**
AFFORDABLE	38	6	7	5
AUTHENTIC	30	6	7	5
TRADITIONAL	58	3	6	2

Comparing the grammatical relations of the selected words, besides the three items that have the same number of relations in the TEC and in the BNC, there is only a slight difference between the number of relations concerning AFFORDABLE and AUTHENTIC. Words that have a fairly general meaning, namely TRADITIONAL and VISIT, are involved in fewer grammatical relations in the TEC, they have double and nearly double relations in the BNC. The investigated words are all keywords in the special corpus that have fairly high frequency of occurrence in the TEC; however, their frequency rates suggest that the number of grammatical patterns is not primarily influenced by frequency. The highly subject-specific word, RESERVATION, has a greater variety of grammatical relations in the specialized TEC than in the BNC.

Looking into the types of relations, the selected nouns form patterns mainly with modifiers, prepositions, adjectives and possessives, and function as object and subject in the TEC, whereas in the BNC, in addition to these patterns, the noun-verb-particle units appear. The verbs occur mainly with modifiers, pronouns, particles and adjectives in the TEC, while in the BNC besides these overlapping patterns they are involved in passive structures, or followed by wh-words. The adjectives occur mainly as modifiers of nouns, corroborating the descriptive nature of tourism discourse, they are constituents of and/or patterns and prepositional phrases, precede and follow verbs, appear in subject + be structures in both corpora, and they are followed by infinitive objects in the BNC. The numbers of grammatical relations suggest that nouns and

verbs tend to form more types of relations than adjectives. The types of the grammatical relations of PASSENGER are illustrated with examples of its grammatical patterns in *Table 31*.

Table 31: *Grammatical relations of key node word PASSENGER with its collocates in the TEC*

PASSENGER	
Grammatical relation	**Example**
modifier of	***economy-class/domestic/average*** *passengers*
nouns modified by	passenger ***plane/flow/comfort***
with verbs as subject	passengers ***complain/left/wore***
with verbs as object	***screen/accommodate/check*** passengers
and/or	***fliers/cases and*** *passengers*
prepositional phrases	***for/with/by*** *passengers*
adjective predicate of	*passengers are* ***prone***
possession	*passenger's* ***final destination***
possessor of	***their*** *passengers*

Scrutinizing *Tables 27* and *28* it seems that the grammatical behaviour of the selected node words is partly influenced by their POS and partly by their meaning, and the fewer types of grammatical relations suggest that collocations are more fixed in the texts of specialized corpora than in those of general corpora. Further investigations on the grammatical and lexical preferences of words can shed light on their collocational behaviour.

6.6.4.3. Grammatical and lexical preferences of collocations

Based on Thurnbury's (2002) categorisation of collocations (as discussed in 3.4.2), to reveal preferences in the colligational relations of the selected node words the following types of grammatical collocations are examined:

- *Noun + to + Infinitive* (N+ T + I) e.g., to cancel, to indulge
- *Infinitive + to +* Verb (I +T + V) e. g. expect to, return to
- *Noun + Preposition* (N + P) e. g. buffet with, view to
- *Preposition + Noun* (P + N) e.g., upon arrival, about delay
- *Auxiliary + Verb* (AUX + V) e.g., should book, might expect
- *Verb + Pronoun* (V + PR) e.g., indulge yourself

To investigate the lexical preferences of collocations the following types are looked at:

- *Adjective + Noun* (ADJ + N) e.g., four-star chef, carry-on baggage

- *Verb + Noun* (V+ N) e.g., serve dishes, visit an attraction, recommend a treatment

- *Noun + Noun* (N + N) e.g., preference of order, distance of travel

- *Adverb + Adjective* (ADV+ADJ) e.g., fully accessible, unbelievably delicious

- *Verb + Adjective* (V +ADJ) e.g., offers affordable, discover authentic

In order to illustrate the collocational patterns and grammatical relations of node words the three randomly selected nouns (N), verbs (V) and adjectives (ADJ) of the full TEC word list were investigated with their right and left binary and four-word span collocates that are included in the collocation lists (see Appendices L, M, and N). The selected node words, their left (L) and right (R) collocates, grammatical and lexical relations are shown in *Table 32*, the grammatical collocations are italicized.

Table 32: Left and right collocates of selected key nodes from the full TEC word list and their grammatical and lexical relations

Part-of-speech	Left collocate	NODE	Right collocate	Grammatical and lexical relations
N	connecting, *cancellations for, itinerary for,* schedule of the	AIRLINE	baggage rules, 's weight allowance, ticket websites	L: V-N, *N-N+P* R: N-N
N	permitted per, *procedure for,* screen, baggage	PASSENGER	check, carrier, comfort, compartment, load, plane	L: ADJ-N, *N-N+P*, V-N, N-N R: N-N
N	make, need, online, *without*	*RESERVATION*	*for*, needed	L: V-N, ADJ-N, *P-N* R: *N-P*, N-ADJ
V	properties, suites, *can*	*BOAST*	*with flat-screen TV/view,* fully furnished	L: N-V, *AUX-V* R: V-P-ADJ-N, V-ADV- ADJ
V	*chance to, time to*	*INDULGE*	with culinary experience food-tasting, *yourself*	L: *N-T-I,* R: V-P-ADJ-N, V-N, *V-PR*
V	definitely, frequently, *want to,* must, never *miss to,* destinations	VISIT	place, island, attraction, beach, temple	L: ADV-V, *I-T-V*, *AUX-V*, N-V R: V-N
ADJ	offers, relatively, more	AFFORDABLE	accommodation, price, deal, rates	L: V-ADJ, ADV-ADJ R: ADJ-N
ADJ	experience, discover, try, taste	AUTHENTIC	food, taste, cuisine, massage, places	L: V-ADJ R: ADJ-N
ADJ	served, flavoured of, enjoy, serve, selection of	TRADITIONAL	environment, massage, dish, cuisine, architecture	L: ADJ-ADJ, V-ADJ, N-ADJ R: ADJ-N

Examining the POS of the target words it is apparent that majority of the left collocates of nouns are adjectives, whereas their right collocates include verbs, nouns, adjectives and prepositions. The left collocates of verbs are mainly auxiliaries and adverbs, while their right collocates are nouns and prepositional structures. The adjectives tend to have more right than left collocates; the left ones are mostly adverbs and verbs, whereas the right ones are generally nouns.

Having a closer look at the colligational and lexical collocational patterns of the selected words, it seems that the examined nouns, *airline*, *passenger*, and *reservation*, tend to have more types of colligational relations with their right collocates than with the left ones; they are either N-P or P-N structures. Only one of the nouns, *reservation*, forms a colligational relation with its left collocate, it is a N-P cluster. The majority of the left and right collocates of the selected nouns form lexical collocational patterns, the most frequent ones with left collocates are V-N and ADJ-N units, whereas N-N and N-ADJ occur mainly as right collocational structures.

The intransitive verb *boast* and *indulge* that can function either as a transitive or intransitive verb, have more right than left grammatical and lexical collocates, whereas the transitive *visit* has more left collocates. Frequent grammatical relations with left collocates include AUX-V, N-T-I and I-T-V patterns, and the only colligational pattern with a right collocate is a V-PR cluster. The relations with right collocates are mostly lexical, longer sequences such as V-P-ADJ-N and V-ADV-ADJ, and V-N units are recurrent patterns.

Interestingly, there are no colligational relations between the selected adjectives, *affordable*, *authentic* and *traditional*, and their left and right collocates. The lexical collocational patterns formed with left collocates include V-ADJ, N-ADJ, ADV-ADJ and ADJ-ADJ units. Reflecting the descriptive nature of tourism language, the typical lexical relation of the adjectives with their right collocates is attributive, they form ADJ-N clusters.

In sum, concerning the grammatical and lexical preference of the target nouns, the types of lexical relations outnumber the colligational ones, yet both left and right collocates are involved in both types of relations. As for verbs, the intransitive ones prefer right collocates that form mainly longer lexical patterns, while the transitive one has more lexical and grammatical relations with its left collocate. The selected adjectives are involved exclusively in lexical relations, mainly with their right collocates.

As this investigation on the grammatical and lexical relations of the selected key node words and their left and right collocates was carried out with only a limited number of words, it does not allow for generalization of the characteristic lexical behaviour and lexicogrammatical patterns of tourism language. However, despite the limited scope of the analysis, it revealed that there is a noticeable diversity of grammatical patterns, and the selected key words tend to have lexical preferences to convey meaning. The analysis of the semantic preferences of key words may disclose what other factors influence their syntactic and lexical behaviour with their collocates.

6.6.5. Semantic preference and semantic prosody

Grounded on the findings of analysis on the grammatical and lexical relations of collocations, the semantic preferences and semantic prosody of selected key words were investigated (semantic prosody and semantic preference are discussed in 2.4.2). Concerning semantic prosody, the majority of words in touristic texts have positive connotation as the ultimate goal is to exert influence on prospective clients. However, as the texts include customer reviews, dissatisfied feedbacks on facilities and services may uncover negative or neutral collocates and semantic prosody, and unveil adverse associative connections as well.

Investigations on the collocational behaviour of a small selection of tourism-specific key node words and their collocates, in addition to providing information on the special terminology and the investigated dominant discursive strategies, may reveal recurrent semantic preferences and give insight into the fixedness of collocates. Meaning is established through consistent sequences of collocates that regularly co-occur, expressing the positive, negative or neutral attitude or evaluation of the speaker/ writer, and the pragmatic meaning acquired by the recipient is referred to as semantic prosody. The closely related concept, semantic preference refers to the semantic categories that are shared by a particular node and its collocates. Although there is a hazy boundary between the concept of semantic prosody and semantic preference, semantic prosody evaluates a topic and indicates to the hearer/reader how to interpret the utterance, whereas semantic preference refers to any semantic category that the collocates are in favour of.

To investigate semantic prosody and semantic preferences in tourism texts 6 key node words were selected that are strongly associated with tourism discourse; some of them may pose

difficulty for tourism students in using them in context. The list of the lexical items includes two nouns: ACCOMMODATION and SERVICE, two verbs: EXPLORE and ENJOY, and two adjectives: EXCELLENT and HISTORIC.

The analysis was conducted with the Thesaurus functionality of *Sketch Engine* software (Kilgarriff & Tugwell, 2001). First, the collocational relations and networks of the target words were scrutinized through the outputs of the Word Sketch function of *Sketch Engine*. As a second step, lists of the 15 most frequent collocates of the target words was created with *Sketch Engine* based on their MI scores, then they were screened to identify semantic relations, finally, their semantic prosody was examined in context in their concordance lines.

To get an insight into the interaction of the semantic preference and the lexical and grammatical structures of the selected key words, it is worth examining the relations of words generated by *Sketch Engine*. *Figure 7* displays the collocational network, the lexicogrammatical and POS relations and semantic preferences of the keyword ACCOMMODATION in a sample *Sketch Engine* output. The size of the circles indicates the frequency of the collocates.

Figure 7: *A sample output of the collocates of the node 'accommodation' generated by Sketch Engine*

In terms of semantic prosody, the majority of the displayed collocates of ACCOMMODATION evoke positive and neutral associations; none of them is inherently negative, which reinforces the persuasive discourse strategy applied in tourism texts. To demonstrate the grammatical relations of the target word the collocates are highlighted with different colours that indicate their POS, and their frequency is exhibited in different sizes. Concerning the grammatical and lexical preferences of this node, it forms lexico-grammatical patterns mainly with verbs, adjectives and nouns that can be grouped into subject-specific semantic sets such as the occupation and description of an accommodation.

A scrutiny of the 15 most frequent collocates of the selected words reveals what semantic patterns are preferred by these items, and whether they have positive, neutral or negative prosody in context. The top 15 collocates of the target words are shown in *Table 33*, generated by *Sketch Engine* Word Sketch functionality. Collocates associated with positive semantic prosody are underlined, words that reflect semantic preference are italicized.

Table 33: The top 15 collocates of the target nodes in Sketch Engine output

ACCOMMODATION	SERVICE	EXPLORE	ENJOY	EXCELLENT	HISTORIC
provide	be	city	*vacation*	service	building
hotel	*excellent*	area	beauty	food	landmark
comfortable	*great*	culture	stay	experience	centre
include	*good*	*visit*	view	restaurant	past
book	offer	*see*	really	massage	site
facility	provide	flora	also	choice	rich
budget	*friendly*	reef	scenery	value	popular
cheap	*attentive*	market	experience	meal	*impressive*
holiday	*professional*	town	*dinner*	wine	*brilliant*
have	customer	beauty	*food*	truly	*superb*
ocean-view	food	experience	*meal*	culinary	cultural
rent	enjoy	alone	treatment	fast	local
affordable	atmosphere	enjoy	thoroughly	*luxurious*	square
reserve	staff	*discover*	relax	tasty	many
recommend	team	safely	*holiday*	local	European

Different semantic relations can be identified among the words including hyponyms of ACCOMMODATION such as *hotel* and *lodging*, and a meronym, *unit*. Semantic relationships of cognitive synonyms can also be detected as *book* and *reserve* can be used interchangeably, and there are near-synonyms like *budget, cheap* and *affordable,* or *excellent, brilliant* and *superb*.

As can be seen in *Table 33*, the proportion of positive collocates is relatively high, more than half of the words evoke pleasant associations, the other words convey fairly neutral prosody, and there is only one word, *alone*, whose positive or negative prosody can be perceived only in context. The considerable number of collocates that are covered by semantic preferences suggests that the collocations of specialised texts tend to be fixed to some extent, and they can be regarded as manifestations of the idiom principle. The converse relationship of *service* and *excellent also* indicates a certain degree of fixedness; *service* is the most frequent collocate of *excellent*, and *excellent* is the second most frequent collocate of *service*. The overlapping words among the collocates including *value, enjoy*, or *experience*, also support the concept that the words become more fixed in specialised texts.

Concerning semantic prosody, a scrutiny of the concordance lines of the target words uncovered that only three target words, ACCOMMODATION, SERVICE and EXPLORE have negative prosody in context.

(a) *Many passengers complain they have to fight or beg for accommodation and eating lunch at an airport on a $5 voucher.*
(b) *Airlines are charging customers for previously free services.*
(c) *We took a full morning to explore the area and still didn't see all the museums, treasures and sights.*

Target nodes ENJOY and EXCELLENT that inherently evoke positive feelings are involved only in collocational patterns that have positive semantic prosody. Nevertheless, HISTORIC exhibits mainly positive and neutral prosody in its concordance lines, negative prosody cannot be associated with this node.

In terms of the semantic preference and the syntagmatic relations of nodes and collocates, besides the meronymy of *hotel* and ACCOMMODATION, the target words tend to favour synonymous collocates. The node ENJOY has two collocates that are substitutable cognitive synonyms, *vacation* and *holiday*; SERVICE attracts near-synonyms like *friendly, attentive* and

professional, while HISTORIC favours synonymous adjectives that have a general heightening effect, including *impressive, brilliant* and *superb*. No instances of antonymy were found between the nodes and collocates or the collocates themselves.

Examining the POS of the 15 most frequent collocates it was found that the target nouns co-occur mainly with verbs, adjectives and nouns, the collocates of verbs are nouns, verbs and adverbs, while adjectives collocate with nouns, adjectives and adverbs. These collocational preferences correspond with the findings on the grammatical and lexical preferences of different target words.

Despite the limitations that there is no established semantic categorisation of collocational preferences and the analysis was carried out with limited items, it provided insight into the semantic preferences and semantic prosody of the target words. The findings on the positive and neutral semantic prosody of tourism key words reaffirms the application of persuasive discourse strategy in touristic texts, and suggest that collocations exhibit a certain degree of fixedness in tourism language.

6.6.6. Considerations of teaching lexical bundles and collocations

6.6.6.1. Theoretical guidelines to teaching collocations

Despite the fact that the advantages of teaching collocations explicitly in the language classroom have been widely confirmed, it still remains a concern how collocations should be taught and what collocations should be given priority. Previous research has provided guidelines; Hill (2001) estimated that up to 70% of spoken and written discourse is made up of multi-word units, and collocations outnumber single-words. Shin and Nation (2008) found that the 1000 most frequent content words of the *BNC* spoken section have 5894 collocates. Additionally, it was discovered that the more frequent the node word, the higher the number of collocates, and the 100 most frequent node words make up nearly 35% of the collocations of the first 1000 node words. These results suggest that teaching explicitly the multiple collocates of the most frequent node words would be beneficial for language learners, and learning gains could be maximized when providing students with wide vocabulary coverage. Nation (2001) maintains that the collocates of the most frequent 2000 words are most valuable to teach explicitly in language classes.

Webb and Kagimoto (2010) found that increasing the number of collocates presented with the node words may increase retention, and it is more effective to learn explicitly several collocates for a small number of node words, than to learn fewer collocates for a larger number of node words. Yet, it is important to consider which node words should be taught with their collocates, Hill (2000) claims that:

'A student with a vocabulary of 2,000 words will only be able to function in a fairly limited way. A different student with 2,000 words, but collocationally competent with those words, will also be far more communicatively competent' (p. 62).

The results suggest that it may be more effective to focus on a smaller number of words, as the learning burden is smaller when more collocates are presented with one node word, and the depth of vocabulary knowledge can be effectively enhanced this way.

6.6.6.2. Pedagogical implications of teaching formulaic language

It was corroborated even in this small-case study that pedagogically-oriented listings of lexical bundles and collocations may help unfold what typical word patterns students need to enhance their tourism vocabulary. As lexical bundles are acquired incrementally like single words, learners need a large number of repeated exposures and practice to become competent in using the appropriate word combinations. Nevertheless, as it was found by Cortes (2004, 2006), lexical bundles are not effortlessly acquired in the natural flow of language learning, thus mere exposure is not enough to achieve the active use of lexical bundles. Noticing, exercises that make learners aware of the nature of lexical bundles and collocations, and activities involving retrieval and practice, such as paraphrasing or substituting lexical items, promote the successful acquisition and retention of multiword sequences. This way, explicit language teaching and tasks that require deep level of language processing foster the acquisition of multiword combinations.

Unfortunately, based on personal experience, in Hungarian language classes teaching words in isolation instead of meaningful context is still in practice, which results in the erroneous transfer of formulaic chunks from students' native language. Since noticing helps transfer the input to intake, with the observation of collocations in authentic texts, including newspaper articles and specialized texts, learners become aware of how real language is used, which can be a great motivating factor. It is also important that phrasal vocabulary should be taught and recorded in an organized way; a beneficial organising principle is relating the collocations to topics and

semantic fields. Students should be encouraged to record new nouns with its collocates, e.g., *a historical/ a contemporary/ a touring exhibition*, and retrieve them with collocation brainstorming exercises.

The TEC word lists and collocation lists can provide guidelines for material and course design in tourism English classes, and can be the source for vocabulary teaching tasks and exercises that focus on the usage of lexical items and lexico-grammatical patterns of tourism language. The collocation lists present the most frequent and relevant collocates of a key node, and give guidance to learners on their grammatical relations and semantic preferences.

6.6.6.3. Classroom activities to raise collocational awareness

Various classroom activities, exercises and strategies can be incorporated into general and ESP lessons to improve the awareness of formulaic language; several activities are suggested by Hill and his co-authors (Hill, Lewis & Lewis, 2000, pp. 98-106) that can enhance students' familiarity with collocations and support learner autonomy:

- On the simplest level, students look for two- or three-word expressions that are related to a topic and underline them.
- Students are given a list of words and they have to find their collocates in a text.
- Students read or listen to a text and have to complete a collocation gap-filling task based on the text.
- Students have to find the odd-one-out collocate of a target word in a list.
- Students have to brainstorm as many collocates of a given word as they can.
- Students have to find the headword that collocates with a list of words.
- Students have to summarise a text using collocations.
- Students have to spot collocational errors in a text.

A set of task types, adapting and customising some of Jablonkai's (2010) suggestions, are presented in Appendix O that illustrate in what ways the findings of the studies on lexical bundles and collocations can be applied in vocabulary teaching.

6.7. Concluding remarks

The analyses in this section aimed at examining three- and four-word lexical bundles in the TEC in terms of their structural and functional categories, as well as identifying collocations

and investigating their grammatical and semantic relations. The application of collocation lists based on self-built corpora is a niche research area. These lists have much potential to exploit in vocabulary teaching; listings of tourism clusters, with the most frequent collocations of key verbs, nouns and adjectives, can be an efficacious means of corpus-based vocabulary extension. Repeated exposure in reading materials, identifying word combinations in concordances, collocational awareness-raising and vocabulary building activities can help enhance in-depth knowledge of the structural patterns and functional uses of lexical bundles, and the grammatical and lexical relations of collocations.

CONCLUSION

The motivation that inspired the writing of the thesis is my deep interest in the language of tourism and corpus linguistics, as well as my several years of teaching experience in ESP, particularly in English in tourism. As there has been modest research on the application of self-built tourism corpora in language teaching, the investigations on the TEC aimed at bridging corpus-based research and pedagogical practices, seeking in what ways the findings can be exploited in teaching the lexis and discourse of tourism. To the analysis of the findings from a pedagogical perspective, research, advances and methodologies in ESP, CL, and vocabulary teaching and learning provided theoretical underpinnings. As corpus-based studies have broadened the scope of linguistic analyses since the 1990s, CL provided methodological framework to the compilation of the TEC and the word- and collocation lists, to the inquiry into the specific lexis and lexico-grammatical patterning of tourism language, as well as into the persuasive promotional features of tourism discourse.

The studies report on lexico-grammatical and discursive analyses of a corpus of tourism industry texts applying corpus-driven and corpus-based research approaches, following Halliday's lexicogrammar approach that regards vocabulary and grammatical structures interdependent. As the TEC comprises extracts of tourism adverts, travel guides, tourist information, travelogues and customer reviews obtained from different websites, they reveal distinct features and patterns of tourism language and discourse including the use of lexis with positive connotations, creating positive imagery with the extensive use of adverbial clusters and intensifiers, or the application of various other persuasive discursive devices and strategies.

The lexico-grammatical patterns were examined at lexical, syntactic and discursive levels, lexical bundles were scrutinized in terms of their grammatical structures and discourse functions. Tourism lexis was investigated in different forms; through contextualising some of the words and word clusters in concordance lines, as well as exemplifying collocational, colligational patterns, semantic and discursive features that are specific to the corpus, thereby to tourism English. The promotional-persuasive discourse function that supports credibility in travel adverts and articles through lexical choices, syntactic structures and discursive techniques, is illustrated with different means of persuasion. The results of the corpus-based and corpus-driven studies instantiate how English language is used in the context of tourism,

raising important and applicable pedagogical implications in teaching vocabulary in ESP classes, specifically in English for tourism purposes.

The studies offer novel contribution to different fields of research; they provide deeper insight into the language of tourism and its specialised vocabulary, reveal several methodological aspects of corpus- and wordlist creation for specific purposes, and shed light on specific discursive patterns and strategies applied in tourism texts. The investigations may help language teachers and ESP practitioners with no or modest experience of corpus linguistics to create a specialized corpus, and compile thematic vocabulary lists or collocation lists for classroom application. The findings of the exploratory studies can provide guidelines to the accomplishment of such pedagogical concerns as how to broaden the breadth and depths of specialized vocabulary, how to encompass specialised core lexis, lexical bundles or collocations, how to identify the discursive features of a special language, and this way raise awareness of context, register and genre.

The compilation of the TEC, thematic word lists and collocation lists contribute to ESP pedagogy at a practical level; as they provide important details on typical lexical items of tourism texts they can form basis for the development of lexical syllabi, teaching materials, annotated vocabulary lists or tourism glossaries, in addition, the applied methodologies are transferable to teaching practice in other fields of ESP.

The analysis of the text coverage of the TEC revealed that the lexis of the selected tourism texts is not highly technical. The lexical items of the word lists fall in the categories that are considered general and sub-technical vocabulary as the texts contain a relatively large proportion of words from the general standard list of the BNC; nearly 88% of the TEC words tally with the first two thousand most frequent words in the BNC.

Investigations on the specific characteristic features the language of tourism shed light on different properties and techniques of the application of linguistic forms and devices that are used to achieve the marketing-orientated goals of tourism. The language of tourism has a discourse of its own that attempts to seduce prospective tourists by influencing their attitude and behaviour through persuasive devices. The persuasive discursive strategies and techniques are manifested in the lexical choices, in the lexicogrammatical patterns and peculiar stylistic features of tourism texts. The high percentage of superlative and comparative forms indicate the tendency of describing the service and attractions only in positive terms, in addition, hyperboles and intensifiers emphasize the uniqueness of the promoted tourism product. Using

the technique of comparison that involves the employment of similes and metaphors, the effect of strangeness and unfamiliarity towards the destination is reduced in terms of spatial and cultural distance. With the technique of creating expectations through the extensive use of adjectives and intensifying adverbial clusters, the promoted tourism entity is presented as the best option to be chosen. Positive reports and testimonies of visitors support the advertised appealing properties of destinations, whereas with the technique of ego-targeting, that is, with the use of pronouns *you* and *we*, the reader is directly addressed and challenged to purchase the touristic experience.

Since insufficient knowledge of the formulaic language and multiword sequences used in tourism discourse may hinder learners of tourism English in expressing their thoughts in appropriate professional manner, teachers of English in tourism need to encourage students to develop an awareness of the collocations they come across during their studies. As many opportunities as possible need to be provided to tourism students to become familiar with word combinations used in tourism texts, and to become cognizant of the distinct, conventionalised features of tourism genre. The scrutiny of the grammatical and lexical preferences of the target words revealed that despite the diversity of grammatical relations, lexical relations are more characteristic to tourism language, tourism keywords tend to bear positive or neutral connotation, and exhibit a certain degree of fixedness. This way the collocational profiles of the keywords presented in the appendices can be the starting point for vocabulary and fluency development activities.

The limitations of the studies and procedures have been discussed in the respective sessions. An obvious overall limitation is that the investigations focus on written tourism texts, further inquiry into spoken communication can reveal distinctive characteristics of spoken tourism discourse. However, data collection for a spoken corpus has to be limited to those domains of tourism language where actual real-life communication takes place with customers, and might raise privacy and business concerns. The other overall limitations, the modest size of the corpus, the limited number of subcorpora and the scantiness of running words, can be eliminated with the extension of the corpus.

The findings of the studies suggest that small-scale self-compiled corpora as the TEC are suitable for linguistic and pedagogical inquiries, and can serve as a basis for material and course design. Further investigations on an extended corpus can unfold more details concerning technical vocabulary, the lexico-grammatical choices in tourism texts, and the lexical and

grammatical behaviour and semantic preferences of collocations. From a pedagogical perspective, grounded on the findings, a longitudinal study on the expansion of tourism students' professional vocabulary and on the development of their discursive competence can be the target of prospective research.

References:

Adolphs, S., & Lin, P.M.S. (2013). Corpus Linguistics. In J. Simpson (Ed.), *The Routledge handbook of applied linguistics* (pp. 597-610). New York: Routledge.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse, *Applied Linguistics*, 24, 425–438.

Aijmer, K. (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Aisenstadt, E. (1979). Collocability restrictions in dictionaries. In *dictionaries and their users*, Exeter Linguistics Studies, (4), R.R.K. Hartman (Ed.), pp. 71-73. Exeter: University of Exeter.

Aston, G. (2000). Corpora and language teaching. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the Third International Conference on Teaching and Language Corpora* (pp. 7-17). Frankfurt am Main: Peter Lang.

Alcantar, C. D. (2007). Construction of a special corpus to extract the most frequent words in a particular field. *Essex Graduate Papers in Language and Linguistics,* 9, 1-17.

Alcantar, C. D. & Scholfield, P. (2010). Reading comprehension and specialized vocabulary: Does tourism vocabulary predict reading comprehension better than general vocabulary? *Language at the University of Essex (LangUE) 2009 Proceedings,* 1-13.

Alcaraz, E. (2000). *El inglés profesional y académico.* Madrid: Alianza.

Alexeeva, L. & Novodranova, V. (2006). A cognitive approach to terminology. In H. Picht (Ed.), *Modern approaches to terminological theories and applications,* pp. 25-34. Bern: Peter Lang.

Allen, J. P., & Widdowson, H. G. (1974). Teaching the communicative use of English. *IRAL: International Review of Applied Linguistics in Language Teaching,* 12 (1), 1-21.

Alonso, A., Millon, C., & Williams, G. (2011). Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci). In I. Kosem & K. Kosem (Eds.) *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011* (pp. 12–22).

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: Research reviews*, (pp. 77-117). Newark, DE.: International Reading Association.

Anderson, J. R. (1990). *Cognitive psychology and its implications: Third edition*. New York: Freeman.

Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics*, 1 (1), 93-111.

Andor, J. (2015). Trains of thought on the issues of syntax, the lexicon and pragmatics: An interview with Noam Chomsky. *International Review of Pragmatics*, 7, 145-155.

Anthony, L. (1997). English for specific purposes: What does it mean? Why is it different? *On-CUE*, 5 (3), 9-10.

Anthony, L. (2011). *AntConc* [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2018). *Introducing English for specific purposes*. New York: Routledge.

Antle, J. B. (2013). Teaching collocations. In N. Sonda & A. Krause (Eds.), *JALT 2012 Conference Proceedings.* Tokyo: JALT.

Ardeo, J. M. G. (2013). (In) compatibility of CLIL and ESP courses at university. *Language Value,* 5 (1), 24-27.

Arno-Macia, E. & Mancho-Bares, G. (2015). The role of content and language in content and language integrated learning (CLIL) at university: Challenges and implications for ESP, *English for Specific Purposes*, 37, 63-73.

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a foreign language*, 4 (2), 91-105.

Baker, M. (1992). *In other words: A coursebook on translation*. London: Routledge.

Barnbrook, G. (2007). Sinclair on collocation. *International Journal of Corpus Linguistics*. 12 (2), 183-199.

Barnbrook, G., Mason, O. & Krishnamurthy, R. (2013). *Collocation: Applications and implications*. Palgrave Macmillan.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Basturkmen, H. (2006). *Ideas and options in English for specific purposes*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Beaver, A. (2012). *A dictionary of travel and tourism*. Oxford: Oxford University Press.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Gilford Press.

Bennet, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbour: University of Michigan Press.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 2, 1-14.

Benson, M., Benson, E., Ilson, R. F., & Young, R. (2009). *BBI Dictionary of English word combinations*. Amsterdam: John Benjamins.

Belcher, D. (2004). Trends in teaching English for specific purposes. *Annual Review of Applied Linguistics*, 24, 165–186.

Belcher, D., Johns, A. M., & Paltridge, B. (2011). Introduction: New directions in English for specific purposes research. *Michigan, English Language Teaching*, 1-4. Accessed on 4th May 2015 at https://www.press.umich.edu/pdf/9780472034604-intro.pdf

Belcher, D., & Lukkarila, L. (2011). Identity in the ESP context. In D. Belcher, A. Johns, & B. Paltridge (Eds.), *New directions for English for specific purposes research* (pp. 73-93). Ann Arbor, MI: University of Michigan Press.

Bernardini, S. (1996). Corpora in the classroom: An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp.15-38). Amsterdam: John Benjamins Publishing Company.

Biber, D. E. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. E. (1993). Representativeness in corpus design. *Literary and Linguistic computing,* 8 (4), 243-257.

Biber, D. E. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In C. Meyer & P. Leistyna (Eds.), *Corpus analysis: Language structure and language use*, pp. 47-70. Amsterdam: Rodopi.

Biber, D. E., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263-286.

Biber, D. E., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25 (3), 371-405.

Biber, D. E., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, D. E, Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, UK: Longman.

Biber, D. E, Reppen, R. & Friginal, E. (2010). Research in corpus linguistics. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics*. pp. 548-570. New York: Oxford University Press.

Biber, D. E. (2012). Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford Handbook of Linguistic Analysis*. pp. 193-224. Oxford: Oxford University Press.

Bijani, H., Rahmani, A., & Asadi, J. (2014). The relationship between strategy use and vocabulary knowledge of the English students. *English for Specific Purposes World*, 43 (15), 1-16.

Bloomfield, L. (1983). *An introduction to the study of language*. Amsterdam: John Benjamins Publishing.

Bojovic, M. (2006). Teaching foreign language for specific purposes: Teacher development. *31st Annual Conference,* Portoroz, Slovenia, 487-493.

Bojovic, M. (2010). Reading skills and reading comprehension in English for specific purposes. *The International Language Conference on the importance of learning professional foreign languages for communication between cultures.* Celje, Slovenia.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum* 1, 1-14.

Booij, G. (2012.) *The grammar of words: An introduction to linguistic morphology*. Oxford: Oxford University Press.

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning ESP: Issues and applications.* pp. 261- 291. Amsterdam: John Benjamins.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (2012). *Corpus-informed research and learning ESP: Issues and applications.* Amsterdam: John Benjamins.

Boulton, A. (2016). Integrating corpus tools and techniques in ESP courses, *ASp la revue de GERAS*, 69, 1-25.

Boulton, A., & De Cock, S. (2017). Dictionaries as aids for language learning. In P. Hanks & G. M. De Schryver (Eds.), *International Handbook of Lexis and Lexicography*. New York: Springer.

Bower, E., Bower, E. R. & Keisler, L. E. (2011). *Building academic language through content-area text: Strategies to support English language learners.* Huntington Beach, CA: Shell Education.

Bowker, L. & Pearson, J. (2002). *Working with specialized language - A practical guide to using corpora.* London: Routledge.

Bracaj, M. (2014). Teaching English for specific purposes and teacher training. *European Scientific Journal,* 10 (2), 40- 49.

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks, *International Journal of Corpus Linguistics*, 20 (2), 139-173.

Brezina, V. & Gablasova, D. (2017). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. *Applied Linguistics*, 38(1), 764-767.

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher*, JALT, Special Issue, 13-16.

Brumfit, J. (1980). *Problems and principles in English teaching*. Principles Institute of English.

Brunet, E. M. M. (2011). English for specific business purposes: English for tourism. In R. G. Moreno, (Ed.), *Words for working: Professional and academic English for international business and economics.* Valencia: Universitat de Valencia.

Bureros, L. L., Tabaranza, Z. L. B. & Roxas, R. R. (2017). Building English - Cebuano tourism parallel corpus and a named - entity list from the Web. *Theory and practice of computation*, 158-169.

Cabré, M. T. (1999). *Terminology: Theory, methods, and applications*. Amsterdam: John Benjamins Publishing.

Calvi, M. V. (2005). *Il linguaggio spagnolo del turismo.* Viareggio: Baroni.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1), 1-47.

Capelli, G. (2006). *Sun, sea, sex and the unspoilt countryside: How the English language makes tourists out of readers.* Pari: Pari Publishing.

Carbajosa, N. (2012). A cognitive experience in ESP: teaching vocabulary to telecommunications engineering students. *ESP World*. Available at http://www.esp-world.info/Articles_2/Experience.html

Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. London: Routledge.

Carter, R. & McCarthy, M. (1996). Correspondence, *ELT Journal*, 50, (4), 369-361.

Carter, R. & McCarthy, M. (2014). *Vocabulary and language teaching*. New York: Routledge.

Carver, D. (1983). Some propositions about ESP, *The ESP Journal*, 2, (2), 131-137.

Case, A. (2013). *The ESP approach: theory and reality of needs analysis and course design.* Accessed on 10th October 2014 at http://www.usingenglish.com/articles/esp-approach-theory-reality-needs-analysis-course-design.html#LrhgTXVFJkOdGwtj.99

Castello, E. (2002). *Tourist-information texts: a corpus-based study of four related genres*. Padova: Unipress.

Castello, E. (2013). Exploring existential and locative constructions in a learner and in an expert corpus of promotional tourist texts. In S. Desoutter, D. Heller, & M. Sala (Eds.), *Corpora is specialized communication*, pp. 385-410. Bergamo: CELSB.

Castello, D. (2014). *A corpus study of strong and powerful*. Master of Arts in Applied Linguistics. University of Birmingham: Birmingham.

Casas-Pedrosa, A. V., Fernandez-Domíngues, J., & Alcaraz-Sintes, A. (2013). Introduction: the use of corpora for language teaching and learning. *Research in Corpus Linguistics*, 1, 1-5.

Chapelle, C.A. (1998). Construct definition and validity inquiry in SLA research. In L.F Bachman, & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research*, pp. 32–70. Cambridge: Cambridge University Press.

Cheng, R. (2014). *John Sinclair's lexical items - an introduction*. Accessed on 30 December 2014 at http://www.commentary.com/lexical_item.html

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Harper & Row.

Chomsky, N. (1986). *Knowledge of language*. New York: Praeger.

Chujo, K., Utijama, M. & Oghigian, K. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System,* 34, (2), 255–269.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language,* 15, (2). Retrieved on 12th November 2014 from http://nflrc.hawaii.edu/rfl/October2003/chung/chung.html

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*, 251-263.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16 (1), 22-29.

Csomay, E. & Petrovic, M. (2012)." Yes, your honour": a corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, 40, 305-315.

Cobb, T. (1999). *The Compleat Lexical Tutor* v.4 accessible at https://www.lextutor.ca/

Cohen, E. (2004). *Contemporary tourism: Diversity and change.* Emerald Group Publishing:
    Bingley, UK.

Collins English Dictionary. https://www.collinsdictionary.com/dictionary/english/vocabulary

Commission of European Communities (2004-2006). *Action Plan.* Available at http://eur-
    lex.europa.eu/legal-content/EN/TXT/HTML/?uri=URISERV:c11068&rid=1
    Amsterdam: Rodopi.

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System*, 27,
    1-18.

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal*, 52 (1), 57-64.

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. Fitzmaurice &
    D. Biber (Eds.), *Using corpora to explore linguistic variation* (Studies in Corpus
    Linguistics, 9), pp. 131-145. Amsterdam: John Benjamins.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples
    from history and biology. *English for Specific Purposes*, 23 (4), 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing
    intensive history class. *Linguistics and Education*, 17 (4), 391-406.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in
    research article introductions. *Journal of English for Specific* Purposes, 12, 33-43.

Cortes, V. (2015). Situating lexical bundles in the formulaic language spectrum: Origins and
    functional analysis development. In V. Cortes & E. Csomay (Eds.), *Corpus-based
    studies in applied linguistics,* pp. 217-220. John Benjamins Publishing.

Council of Europe (2001). *CEFR (Common European Framework of Reference for Languages)
    manual.* Strasbourg: Language Policy Division. Accessed on 5th May 2015 at
    http://www.coe.int/t/dg4/linguistic/Source/Framework_EN. pdf

Cowie, A. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied
    Linguistics, (*2), 223-235.

Coxhead, A. (2002). A new academic word list. *TESOL Quarterly*, 34, 213-238.

Coxhead, A. (2012). Specialised vocabulary in secondary school classrooms: Teachers' perspectives. *RELC Anthologies*.

Coxhead, A. (2013). Vocabulary and ESP. In B. Paltridge & S. Starfield (Eds.). *The Handbook of English for Specific Purposes,* pp. 115-132. Boston: Wiley-Blackwell.

Coxhead, A. (2017). *Vocabulary and English for specific purposes research*: *Quantitative and qualitative perspectives*. New York: Routledge.

Crandall, J. (1992). Content-centered learning in the United States. *Annual Review of Applied Linguistics*, 13, 111-126.

Crandall, J. (1994). *Content-centered language learning*. ERIC Clearinghouse on Languages and Linguistics: Washington DC. Accessed 06/05/2015 at http://www.ericdigests.org/1994/content.htm.

Cresswell, A. (2007). Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the Foreign Language Classroom* (pp. 267-288). Amsterdam: Rodopi.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics* 34, 369–88.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework*, pp. 3-49. Los Angeles: California State University, Evaluation, Dissemination, and Assessment Center.

Cummins, J. (2002). Reading and the bilingual student: Face and friction. In G.G. Garcia (Ed.), *English learners: Reaching the highest level of literacy,* pp. 2-33. Newark, DE: International Reading Association Inc.

Dann, G.M.S. (1996). *The language of tourism: a sociolinguistic perspective*. Wallingford: CAB International.

Davies, B. G. (1993). *Tools for teaching*. San Francisco: Jossey-Bass.

Davies, S. (2003). Content based instruction in EFL context. *The Internet TESOL Journal*, 9 (2). Available at http://iteslj.org/Articles/Davies-CBI.html

Day, J. & Krzanowski, M. (2010). *English for specific purposes: an introduction.* Cambridge: Cambridge University Press.

Deardorff, D. K. (2006). Identification and assessment of intercultural competence as a student outcome of internationalization. *Journal of Studies in International Education*, 10, (3), 241–266.

DeCock, S. (2003). *Recurrent sequences of words in native speaker and advanced learner spoken and written English: a corpus-driven approach*. Unpublished PhD thesis. Louvain-la-Neuve: Université catholique de Louvain

DeMarco, C. (2011). The role of register analysis in an English for special purposes (ESP) curriculum – Accessed on 20th April 2015 at: http://www.tesol.org/news-landing-page/2011/10/31/the-role-of-register-analysis-in-an-english-for-special-purposes-%28esp%29-curriculum-%28from-winter-1986-vol.-2-no.-2%29#sthash.PbOFgh4o.dpuf

Dickinson, M. (2013). *Corpus linguistics: Application 2: Collocations*. Linguistics Department Indiana University. Retrieved on 30/12/2014 from http://cl.indiana.edu/~md7/13/615/apps/02-collocations/02-collocations-2x3.pdf

Dóczi, B. (2011). Comparing the vocabulary learning strategies of high-school and university students: A pilot study. *Working Papers in Language Pedagogy (WoPaLP)*, 5, 138-158.

Donesch-Jezo, E. (2013). Using language corpus in teaching foreign language vocabulary. *International Interdisciplinary e-Journal.* 2 (1), 11-25.

Dörnyei, Z. (2005). *The pshychology of the language learner: Individual differences in second language acquisition.* Mahwah, New Jersey: Earlbaum Associates Publishers.

Dubicka, I. & O'Keefe, M. (2003). *English for international tourism*. Essex: Longman.

Dudley-Evans, T. & St. John, M. (1998). *Developments in ESP: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Dudley-Evans, T. (2000). Genre analysis: a key to the theory of ESP? *Ibérica,* 2, 3-11.

Duenas, M. (2004). The whats, whys and whos of content-based instruction second/foreign language education. *International Journal of English Studies,* 4 (1), 73-96.

Dugan, C. (2010). *Strategies for building academic vocabulary in social studies*. Huntington Beach, USA: Educational Publishing.

Durrant, P. (2008). *High-Frequency collocations and second language learning.* (Unpublished doctoral dissertation). Nottingham, UK: University of Nottingham.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28, 157-169.

Durrant, P. & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory,* 6 (2), 125-155.

Durrant, P. (2014). Discipline- and level-specificity in university students' written vocabulary. *Applied Linguistics*, 35(3), 328-356.

Echevarria, J., Vogt, M. & Short, D. J. (2004). *Making content comprehensible for English learners: The SIOP model.* Boston, MA: Pearson Education Inc.

Edwards, P. & Curado, A. (2003). The promotion of tourism through key concepts and specific concepts. *LSP & Professional Communication*, 3(1), 26- 41.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91- 126.

Ellis, N. C. (1997). Vocabulary acquisition: word structure, collocation, word-class, and meaning. In N. Schmitt, & M. J. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp.122-139). Cambridge: Cambridge University Press.

Ellis, N. C., Simson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly*, 42 (3), 375-396.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal Teddy bear. *Annual Review of Applied Linguistics*, 32, 17-44.

English-Hindi Tourism Text Corpus   http://tdil-
dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1418&lang=en

Eurydice European Unit. (2006). *Content and language integrated learning (CLIL) at school in Europe.* Brussels: Eurydice European Unit.

Ewer, J.R. & Latorre, G. (1969). *A course in basic scientific English.* London: Longman.

Faber, P. B., & Usón, R. M. (1999). *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter.

Felder, R., & Brent, R. (1996). Navigating the bumpy road to student-centered instruction. *College Teaching*, 44(2), 43-47.

Fiorito, L. (2005). *Teaching English for specific purposes (ESP).* Accessed on 17/05/ 2014 at http://www.usingenglish.com/articles/teaching-english-for-specific-purposes-esp.html

Firth, J. R. (1957). *Papers in Linguistics. 1934-1951*. In U. Römer (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts, and didactics*. Amsterdam: John Benjamins Publishing.

Flowerdew, J. (1992). Definitions in science lectures. *Applied Linguistics*, 13(2), 202-221.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System*, 21(2), 231-244.

Flowerdew, J. (2012). English for research publication purposes. In B. Paltridge & S. Starfield (Eds.). *The handbook of English for specific purposes*. John Wiley & Sons.

Flowerdew, L. (2004). The argument for using English specialised corpora to understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions; Perspectives from corpus linguistics*, (pp. 11-36). Amsterdam: John Benjamins Publishing.

Flowerdew, L. (2009). Applying corpus linguistics to language pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, 14 (3), 393-417.

Flowerdew, L. (2012). *Corpora and language education*. Basingstoke: Palgrave: McMillan.

Folse, K. (2010). Is explicit vocabulary focus the reading teacher's job? *Reading in a Foreign Language*, *22*, 139-160.

Fox, R. (2008). English in tourism: A sociolinguistic perspective. *Tourism and Hospitality Management*, 14 (1), 13-22.

Francesconi, S. (2014). *Reading tourism texts: A multimodal analysis.* Bristol: Channel view Publications.

Fraser, S. (2005). The lexical characteristics of technical texts. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.), *JALT2004 conference proceedings* (pp. 318-327). Tokyo, Japan: Japan Association for Language Teaching.

Fuentes, A. C. (2004). The use of corpora and IT in evaluating oral task competence for tourism English. *CALICO Journal*, 22 (1), 5-22.

Fuentes, A. C. (2007). A corpus-based assessment of reading comprehension in English. In E. Hidalgo, L. Quereda, & J. Santana (Eds), *Corpora in the Foreign Language Classroom*, (pp. 309-326). Amsterdam: Rodopi.

Fuentes, A. C. (2015). *A corpus-based focus on ESP teaching*. Accessed on 30/01/2019 https://www.researchgate.net/publication/268260420

Fujita, R. (2009). Using an ad hoc corpus to create an ESP vocabulary list for students majoring in tourism. Paper presented at the *JACET 48th Annual Convention Program*, Hokkai-Gakuen University.

Fujita, R. (2010). Generating English vocabulary lists for students majoring in tourism. *JACET Journal*, 51, 1-13.

Fuster-Márquez, M. (2014.) Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction*, 7 (1), 84 – 121.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *TESL-EJ 8* (4). https://files.eric.ed.gov/fulltext/EJ1068106.pdf

Gabrielatos, C. (2017). Clusters of Keyness: A principled approach to selecting key items. *Corpus Linguistics in the South 15*, University of Cambridge, 28 October 2017.

Gabrielatos, C. (2018). The lexicogrammar of BE interested: description and pedagogy. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L.M. Dillmann (Eds.), Corpora and Lexis. *Language and Computers*, 81, 240-276.

Gairns, R. & Redman, S. (1986). *Working with words*. Cambridge: Cambridge University Press.

Gandin, S. (2013). Translating the language of tourism. A corpus based study on the Translational Tourism English Corpus (T-TourEC). *Procedia - Social and Behavioral Sciences*, 95, 325 – 335.

Gandin, S. (2014). Investigating loan words and expressions in tourism discourse: A corpus-driven analysis on the BBC-travel corpus. *European Scientific Journal,* 10 (2), 1-17.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, 28(2), 241-265.

Gardner, D. (2013). *Exploring vocabulary: Language in action.* New York: Routledge.

Gardner, D. & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics*, 35 (3), 305- 327.

Gass & L. Selinker (2008). *Second language acquisition: An introductory course.* New York: Routledge/Taylor Francis.

Gatehouse, K. (2001). Key issues in English for specific purposes (ESP) curriculum. *The Internet TESL Journal*, *7*, 1-11.

Gavioli, L. (2002). Some thoughts on the problem of representing ESP through small corpora. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 293–303). Amsterdam: Rodopi.

Gavioli, L. (2005). *Exploring corpora for ESP learning.* Amsterdam: John Benjamins.

Gee, J. P. (2005). *An introduction to discourse analysis: Theory and method*. London: Routledge.

Gee, J. P. (2011). *How to do Discourse Analysis: A Toolkit*. New York: Routledge.

Genesee, F. (1994). *Integrating language and content: Lessons from immersion*. Educational Practice Report 11, National Center for Research on Cultural Diversity and Second Language Learning.

Georgieva, V. (2010). *Teaching terms effectively in ESP*. Retrieved 12/11/2014 from www.beta-iatefl.org/.../2010-Valentina-Georgieva.doc

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning, *Language Teaching*, 40, (2), 87-118.

Gledhill, C. J. (2000). *Collocations in science writing*. Gunter Narr Verlag.

Gonda, T. & Csapó J. (2019). The travel habits of the Hungarian population and the sources of the information affecting their travel decisions. In T. Gonda, J. Mátyás, Z. Raffay & J. Csapó (Eds.), *Turizmus, fogyasztás, generációk*. II. Nemzetközi Turizmusmarketing Konferencia: Absztraktkötet. (pp. 107-108). Pécs: Pécsi Tudományegyetem Közgazdaságtudományi Kar.

Gotti, M. (2003). *Specialized discourse: linguistic features and changing conventions*. Bern: Peter Lang.

Gotti, M. (2006). The language of tourism as specialized discourse". In O. Palusci & S. Francesconi (Eds.), *Translating tourism: Linguistic/cultural representations*. (pp. 15-34). Trento: Editrice Università degli Studi di Trento.

Grabe, W., & Stoller, F. L. (1997). Content-based instruction: Research foundations. In M.A. Snow & D. M. Brinton, (Eds.), *The content-based classroom: Perspectives on integrating language and content,* pp. 5-21. White Plains, NY: Longman.

Granger, S. (2008). Learner corpora. In A. Lüdeling, & M. Kytö, (Eds.), *Corpus Linguistics: An International Handbook,* pp. 259-275. Berlin & New York: Walter de Gruyter.

Granger, S. & Paquot, M. (2010). *Lexicography in the 21st century: New challenges, new applications*. Proceedings of ELEX2009. Cahiers du CENTAL N°7, 87-96. Louvain-la-Neuve, Presses universitaires de Louvain.

Granger, S. (2011). From phraseology to pedagogy: Challenges and prospects. In T. Herbst, S. Schüller & P. Uhrig (Eds.), *The Phraseological view of language. A tribute to John Sinclair*. pp. 123-146. Berlin & New York: Mouton de Gruyter.

Granger, S. & Paquot, M. (2012). *Electronic lexicography*. Oxford: Oxford University Press.

Graves, M. F. (2006). *The vocabulary book: Learning and instruction*. Teachers College Press.

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, pp. 3-25. Amsterdam: John Benjamins.

Gries, S. Th. (2013). *Statistics for linguistics with R: A practical introduction*. Boston: Walter de Gruyter.

Griffin, G., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, 17, 443-460.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. Lund: Lund University.

Habil, H., Hassan, H., & Nasir, Z. M. (2008). Persuasive strategies of tourism discourse. In H. Habil & H. Hassan (Eds.), *New perspectives in language and communication research*, pp. 1-20. Malaysia: Univision Press.

Hadley, J. (2006). Needs analysis in ESP. In M. S. Lahlou, & A. Richardson, (Eds.), *English for specific purposes in the Arab world*, pp. 3-6. Dubai: TESOL Arabia.

Halliday, M.A.K. (1980). "Register variation", *Sophia Linguistica* 6, 60-74.

Halliday, M. A. K. (1988). On the language of physical science, In M. Ghadessy (Ed), *Registers of written English: Situational factors and linguistic features,* pp.162-178. London: Pinter.

Halliday, M. A. K. (1991). *Corpus studies and probabilistic grammar*. Routledge.

Halliday, M.A.K., McIntosh, A. & Strevens, P. (1964). Exploring Corpora for ESP Learning: The linguistic sciences and language teaching. London: Longmans.

Handl, S. (2008). Essential collocations for learners of English: The role of collocational direction and weight, In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language teaching and learning*, pp.43-66. Amsterdam: John Benjamins.

Hanus, F. & Fürsich, E. (2014): On the relevance of travel journalism: an introduction. In: F. Hanus & E. Fürsich (Eds.), *Travel journalism: exploring production, impact and culture*, pp. 1-18. London: Palgrave McMillan.

Harding, K. (2007). *English for specific purposes: Resource books for teachers*. Oxford: Oxford University Press.

Harris, R. & Howard, J. (1996). *Dictionary of travel, tourism and hospitality terms*. Hospitality Press.

Havril, A. G. (2009). *A szaknyelv térhódítása a felsőoktatásban.* (Unpublished PhD Dissertation). Veszprém: Pannon University.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21, 303-317.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development - a progress report. *EUROSLA Monograph Series*, 2, 29-56.

Hiebert, E. H. & Kamil, L. M. (2005). *Teaching and learning vocabulary: Bringing research to practice.* Mahwah, New Jersey: Erlbaum Associates.

Hill, J. (1999). Collocational competence. *English Teaching Professional*, 11, 3-6.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In Lewis, M. (Ed.), *Teaching collocation: Further developments in the lexical approach*, pp. 47-69. Hove: Language Teaching Publications.

Hill, J., Lewis, Morgan & Lewis, Michael (2000). Classroom strategies, activities and exercises. In Lewis, M. (Ed.), *Teaching Collocation Further developments in the lexical approach*, pp. 88-117. Hove: Language Teaching Publications

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Hoey, M. (2000). A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocation*. Hove: Language Teaching Publications.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London, UK: Routledge.

Hoey, M. (2009). Lexical priming. *MED Magazine*. Accessed on 31st January 2014 at http://www.macmillandictionaries.com/MED-Magazine/January2009/52-LA-LexicalPriming.htm

Holmes, W., Bialik, M., & Fadel, C. (2023). *Artificial intelligence in education: Promises and implications for teaching and learning*. PDF downloaded from ucl.ac.uk on 4th July 2023

Horváth, J. (2001). *Advanced writing in English as a foreign language: A corpus-based study of processes and products*. Pécs: Lingua Franca Csoport.

Horváth, J. (2006). Keywords in context: Corpus linguistics, In B. Richter (Ed.), *First steps in theoretical and applied linguistics* (pp. 111-117). Budapest: Bölcsész Konzorcium.

Hou, H. I. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching*, 7(5), 26-35.

Howarth P. (1998): Phraseology and Second Language Proficiency, *Applied Linguistics*, 19, 24-44.

Hu, M. H-C., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13 (1), 403-430.

Huang, P. (2010). Where are we ESP practitioners? A report of the First International Conference "ESP in Asia." *Asian ESP Journal,* 8-19.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). London: Macmillan.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge: Cambridge University Press.

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* John Benjamins Publishing.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centered approach.* Cambridge: Cambridge University Press.

Hüttner, J., Smit, U., & Mehlmauer-Larcher, B. (2009). ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis. *ScienceDirect*, System 37 (2009) 99-109.

Hyland, K. (2006). English for specific purposes: some influences and impacts. In A. Cummins & C. Davison (Eds.), *International handbook of language teaching* (pp. 379-390). Springer Science & Business Media.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation, *ScienceDirect*, 27 (1), 4-21.

Hyland, K. (2011). Specific purpose programs. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching*. John Wiley & Sons. Google eBook available at https://books.google.hu/books?id=h3n_mZtVTWoC&printsec=frontcover&vq=specific+purposes&source=gbs_ge_summary_r&cad=0#v=onepage&q=specific%20purposes&f=false

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), S*ociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Jablonkai, R. R. (2010). *A corpus-linguistic investigation into the lexis of written English EU discourse: An ESP pedagogic perspective*. (Unpublished doctoral dissertation). Budapest: Eötvös Lóránd University.

Jablonkai, R. R. & Cebron, N. (2017). Corpora as tools for self-driven learning: A corpus-based ESP course. In N. A. Alias & J. E. Lauran (Eds.) *Student-driven learning strategies for the 21st century classroom* (pp. 274-298). Accessible at https://www.igi-global.com/chapter/corpora-as-tools-for-self-driven-learning/171583?camid=4v1

Jalilzadeh, K., & Tahmasebi, A. (2014). Content-based syllabus, *European Scientific Journal,* 1, 223-231.

Jarema, G. & Libben, G. (2007). *The mental lexicon: Core perspectives*. Bingley, UK: Emerald Group Publishing.

Javid, C. Z. (2013). English for specific purposes: Its definitions, characteristics, scope and purpose. *European Journal of Scientific Research*, 112, (1), 138-151.

Jaworski, A., & Thurlow, C. (2010). Language and the globalizing habitus of tourism: Towards a sociolinguistics of fleeting relationships. In N. Coupland (Ed.), *Handbook of Language and Globalisation.* (pp. 255–286). Oxford: Wiley-Blackwell.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns, & P. King (Eds.), Classroom concordancing. *ELR Journal*, 4, 27-45.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In: B. Kettermann, & G. Marko (Eds.), *Teaching and learning by doing corpus linguistics*, (pp.107-117). Amsterdam: Rodopi.

Johns, A. & Makalela, L. (2011). Needs analysis, critical ethnography, and context: Perspectives from the client and the consultant. In D. Belcher, A. M. Johns & B. Paltridge (Eds.), New directions in English for specific purposes research, (pp. 197-221). Ann Arbor: The University of Michigan Press.

Johns, A. M., & Salmani-Nodoushan, M. A. (2015). English for specific purposes: The state of the art. (An online interview with Ann M. Johns), *International Journal of Language Studies*, 9, (29), 113-120.

Jones, L. (1998). *Welcome! English for international tourism*. Cambridge: Cambridge University Press.

Kang, N. & Yu, Q. (2011). Corpus-based stylistic analysis of tourism English. *Journal of Language Teaching and Research*, Vol. 2, No. 1, 129-136.

Károly, A. (2005). The importance of raising collocational awareness in the vocabulary development of intermediate level learners of English. *Eger Journal of English Studies*, 5, 58-69.

Keck, C. (2013). Corpus linguistics in language teaching. In C. A. Chapelle (Ed.), *The Encyclopaedia of Applied Linguistics*, (pp.1-3). Blackwell Publishing.

Kennedy, G. (1990). Collocations: where grammar and vocabulary teaching meet. In S. Anivan (Ed.), Language Teaching Methodology for the Nineties. *RELC Journal,* Singapore: 215-229.

Kennedy, G. (1998). *An introduction to corpus linguistics.* New York: Longman.

Kilgariff, A. & Rundell, M. (2002). Lexical profiling software and its lexicographic applications – A case study. *EURALEX 2002 Proceedings*, pp. 807-818.

Kilgariff, A. & Tugwell, D. *Sketch Engine* corpus software. Available at https://www.sketchengine.eu/

Kiss, I. (2012). A special journey: Exploring self-compiled corpora in vocabulary teaching. (Unpublished MA thesis)

Kiss, I. & Horváth, J. (2015). Sheltered Beaches: A Tourism Collocation Approach to CLIL Vocabulary Teaching. In M. Lehmann, R. Lugossy, & J. Horváth (Eds.), *UPRT 2015 Empirical Studies in English Applied Linguistics* (pp. 166-178). Pécs: Lingua Franca.

Knight, K. (2012). Teaching business case studies through power point presentations*, The Journal of Kanda University of International Studies*, 24, 1- 28.

Knight, K., Lomperis, A., van Naerssen, M., & Westerfield, K. (2010). English for Specific Purposes: An overview for practitioners and clients, *The Journal of Kanda University of International Studies*, 24, 1- 22.

Knowles, G. & Don, Z. M. (2004). The notion of a 'lemma': headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9 (1), 69-81.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). New York: Routledge.

Korosadowicz-Struzynska, M. (1980). Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensia*, 12, 109-120.

Kosem, I. & Krishnamurthy, R. (2007). *A new venture in corpus-based lexicography: towards a dictionary of academic English*. Accessed 24/02/2019 at https://www.academia.edu/13010391/A_new_venture_in_corpus-based_lexicography_towards_a_dictionary_of_academic_English

Krahnke, K. (1987). *Approaches to syllabus design for foreign language teaching*. New York: Prentice Hall.

Krishnamurthy, R. (2002). *Corpus, collocation and lexical sets*. Accessed 24/02/2019 at https://www.researchgate.net/publication/42363734_2002-Corpus_collocation_and_lexical_sets

Kroll, J. F., & De Groot A. M. B. (Eds.). (2005). *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press.

Kwary, D. A. (2011). A hybrid method for determining technical vocabulary. *System*, 39, 175-185.

Laborda, J. G. (2009). Using webquests for oral communication in English as a foreign language for Tourism Studies. *Educational Technology & Society*, 12 (1), 258–270.

Lam, P.Y.W. (2004). A corpus-driven lexico-grammatical analysis of English tourism industry texts and the study of its pedagogical implications in English for specific purposes. In *Corpora in the foreign language classroom: selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*, University of Granada, Spain, 4-7 July, 2004. Encarnación H. Tenorio, Luis Q. Rodríguez-Navarro, Juan, Santana (Eds.), (2007). Rodopi.

Larsen-Freeman, D. (2000). *Techniques and principles in language teaching.* Oxford: Oxford University Press.

Laufer, B. (1997). What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In N. Schmitt, and M. J. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.140-155). Cambridge: Cambridge University Press.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, 28(1), 89-108.

Laufer, B., & Paribakht, S. (1998). The relationship between active and passive vocabularies: Effects of language learning context. *Language Learning*, 48, 365-391.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15-30.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, & B. Altenberg (Eds.), (1992). *English corpus linguistics: studies in honour of Jan Svartvik* (pp.8-29). London: Longman.

Leech, G. (1997/a). Introducing corpus annotation. R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp.1-18). London: Longman.

Leech, G. (1997/b). Teaching and language corpora. A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), Teaching language and corpora. (pp. 1-23). New York: Addison Wesley, Longman.

Lehmann, M. (2009). *Assessing English majors' vocabulary at the University of Pécs: Working towards a corpus-based approach*. (Unpublished PhD dissertation.)

Lei, H. (2021). High school students' foreign language vocabulary acquisition in the era of artificial intelligence. *Advances in Social Science, Education and Humanities Research*, 637, 659-662.

Lesniewska, J. (2006). Collocations and second language use. *Studia Linguistica*, pp. 95-104. Retrieved 24/02 /2014 from http://www.filg.uj.edu.pl/documents/41616/4333138/12306-Lesniewska.pdf.

Lessard-Clouston, M. (2012/2013). Word lists for vocabulary learning and teaching. The CATESOL Journal, 24 (1), 287-304.

Lewis, M. (1993). T*he lexical approach: The state of ELT and a way forward.* Hove, UK: Language Teaching Publications.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice.* Hove: Language Teaching Publications.

Liu Na. & Nation, P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16, 35-42.

Liuolienė, A. & Metiunienė, R. (2013). Creativity in building ESP vocabulary in the context of ICT. *SANTALKA: Filologija, Edukologija, 21(1), 41-52.*

Long, M. (2005). *Second language needs analysis*. Cambridge, UK: Cambridge University Press.

Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffner, H. Sauntson, & G. Fox (Eds.), *Words in context: A tribute to John Sinclair on his retirement.* Birmingham: University of Birmingham.

Lowe, (I). (2010). *Authentic lesson material in ESP*. Accessed 04/05/2015 at http://www.scientificlanguage.com/esp/authentic.pdf

Luka, I. (2004), Development of Students' ESP Competence in Tertiary, Paper presented at the International bilingual conference "*Assessing language and (inter-)cultural competences in Higher Education*" in Finland, the University of Turku, 30th-31st August, 2007. Retrieved 25/04/2015 from http://www.mig-komm.eu/system/files/Language_competence.pdf

Luka, I. (2007). *Development of students' ESP competence in tertiary studies*. Paper presented at the International bilingual conference "Assessing language and inter-cultural competences in Higher Education" in Finland, the University of Turku, 30.-31 August, 2007. Retrieved from http://www.mig-komm.eu/system/files/Language_competence.pdf

Lyons, J. (1977). *Semantics.* Cambridge University Press.

Ma, Q. (2009). *Second language vocabulary acquisition*. Bern: Peter Lang.

Maci, S. (2010). *The language of tourism*. Bergamo: CELSB.

Maci, S. (2012). Tourism as a specialised discourse: The case of normative guidelines in the European Union. *Token: A Journal of English Linguistics.* 1, 37-58.

Mackay, R., & Mountford, A. (1978). *English for specific purposes*. London: Longman.

Madrid, D. & García Sánchez, E. (2001). Content-based Second Language Teaching. In E. García Sánchez (Ed.), *Present and Future Trends in TEFL* (pp. 101-134)*. Universidad de Almería: Secretariado de publicaciones.

Maican, M. A. (2014). Teaching English to tourism students. *Bulletin of the Transilvania University of Brasov*, Series V, 7, (56) 2.

Manca, E. (2019). *Persuasion in tourism discourse: Methodologies and models*. Cambridge: Cambridge Scholars Publishing.

Marco, M. J. L. (2002). Internet content-based activities for ESP. *English Teaching Forum*, 40, (2), 20-25.

Marzá, N. E. (2009). *The specialised lexicographical approach: A step-further in dictionary making.* Germany: Peter Lang.

Marzano, R. J. (2004). *Building background knowledge for academic achievement: Research on what works in schools.* Alexandria, VA, USA: Association for Supervision and Curriculum Development.

Marzano, R. J. & Pickering, (2006). *Building academic vocabulary: Teacher's manual*. Association for Supervision & Curriculum Development.

Master, P. (1997). Content-based instruction vs. ESP. *TESOL Matters*, 7 (6), 10-11.

Mayo, M. P. G. (2007). *Investigating tasks in formal language*. Multilingual Matters.

McCarthy, M. & O'Dell, F. (2005). *English collocations in use*. Cambridge: Cambridge University Press.

McEnery, T., & Hardy, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. & Yukio, T. (2006). *Corpus-based language studies: An advanced resource book.* London: Routledge.

Meyer, C. F. (2004). English corpus linguistics: An introduction. Cambridge: Cambridge University Press.

McIntosh, C., Francis, B., & Poole, R. (2009). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*. doi:10.1093/applin/amw050

McMillan (2010). *Macmillan Collocations Dictionary*.

Meara, P. & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grundwell (Ed.), *Applied linguistics in society,* (pp. 80-87). London: CILT.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition*, (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt, and M. J. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109-121). Cambridge: Cambridge University Press.

Medlik, S. (2003). *Dictionary of travel, tourism and hospitality*. Butterworth Heinemann.

Mel'cuk, I. (1981). Meaning-text models: A recent trend in Soviet linguistics. *Annual Review of Anthropology*. 10, 27-62.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt, & M. J. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Meuter, R. (2009). Neurolinguistic contributions to understanding the bilingual lexicon. In A. Pavlenko (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches*. Bristol, UK: Multilingual Matters. pp. 1-25.

Meyer, C. F. (2004). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Monterrubio, J. C. (2011). *Tourism and sociocultural change: A conceptual perspective*. Mexico: Plaza y Valdez y Universidad Autónoma de México.

Moreno, R. G. (2011). *Words for working: Professional and academic English for international business and economics.* Valencia: Universitat de Valencia.

Motuziene, R. (2013). A task-based language teaching approach in teaching ESP. *Cypriot Journal of Education Sciences*, 8 (3), 331-340.

Mukoroli, J. (2011). Effective vocabulary teaching strategies for the English for Academic Purposes ESL classroom. *MA TESOL Collection,* 1 (1), 1-47.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233–253.

Nagy, W. E., & Scott, J. A. (2000). Vocabulary processes. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (pp. 269–284). Mahwah, NJ: Erlbaum.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I.S.P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge, UK: Cambridge University Press.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63, 1, 59-82.

Nation, I.S.P. (2007). The four strands. *Innovation in Language Learning and Teaching*, 1(1), 1-12.

Nation, I.S.P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.

Nation, I.S.P. (2016). *Making and using word lists for language learning and testing.* Amsterdam: John Benjamins Publishing.

Nattinger, J. R. & DeCarrico J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Navracsics, J. (2007). Word classes and the bilingual mental lexicon. In Zs. Lengyel & J. Navracsics, (Eds.), *Second language lexical processes: Applied linguistic and psycholinguistic perspectives.* Clevedon, UK: Multilingual matters.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Nesselhauf, N. (2011). *Corpus linguistics: A practical introduction*. Accessed 16/02/2019 http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Nigro, M. G. (2006). The language of tourism as LSP? In H. Picht (Ed.), *Modern approaches to terminological theories and applications* (pp. 185-198). Germany: Peter Lang.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language use and Language Teaching.* Cambridge: Cambridge University Press.

Orr, T. (2002). *English for specific purposes*. Alexandria, Virginia: TESOL.

Oxford Latin Dictionary (1968). Accessed on 30th January 2019 at https://archive.org/details/OxfordLatinDictionary_201708

Oxford Advanced Learner Dictionary (2020). Accessed on 1st February 2020 at https://www.oxfordlearnersdictionaries.com/

Pace-Sigge, M. T. L. (2013). The concept of lexical priming in the context of language use. *ICAME Journal*, 37, 149-173.

Paquot, M. (2015). Lexicography and phraseology. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of corpus linguistics*. Cambridge: Cambridge University Press.

Pastor, G. S. (2017). Collocational constructions in translated Spanish: What corpora reveal. In R. Mitkov (Ed.), *Computational and corpus-based phraseology: Second International Conference, Europhrase*, (pp. 29-40). London, UK: Springer.

Pérez-Paredes, P. (2020). *Corpus Linguistics for education: A guide for research*. Routledge.

Peters, P. & Fernandez, T. (2013). The lexical needs of ESP students in a professional field. *English for Specific Purposes*, 32, 236-247.

Peterwagner, R. (2005). *What is the matter with communicative competence?: An analysis to encourage teachers of English to assess the very basis of their teaching.* Wien: Lit Verlag.

Pica, T. (2002). Subject-matter content: How does it assist the interactional and linguistic needs of classroom language learners? *Modern Language Journal*, 86 (1), 1-19.

Pierini, P. (2009). Adjectives of tourism English on the web: A corpus-based study. *CÍRCULO de Lingüística Aplicada a la Comunicación*, 40, 93-116.

Pitts, M., White, H. & Krashen, S. (1989). Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 5 (2), 271-275.

Pojanapunya, P. & Watson Todd, R. (2016) Log-likelihood and odds ratio: keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14 (1).

Range vocabulary analysis program, available at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Rea, R. C. (2009). A first approach to the lexical profile of telecommunication English: frequency, distribution, restriction and keyness. *Revista de Linguistica y Lenguas Aplicadas*, 4, 161-173.

Read, J. (2004/a). Plumbing the depth: How should the construct of vocabulary knowledge be defined? In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing*. (pp. 209-227). Amsterdam: John Benjamins Publishing.

Read, J. (2004/b). Research in teaching vocabulary, *Annual Review of Applied Linguistics*, 24, 146-161.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *IJES*, 7 (2), 105-125.

Rébék-Nagy, G. (2013). A szaknyelvoktatás aktuális kérdései. Szakmai workshop prezentáció, Szombathely. www.slideserve.com/newman/dr-r-b-k-nagy-g-bor-pte-ok-nyelvi-int-zet-szaknyelvoktat-k-s-kutat-k-orsz-gos-egyes-lete-szokoe/

Reppen, R. (2010). Building a corpus: what are the key considerations? In A. O'Keefe, & M. McCarthy (Eds.), The *Routledge handbook of corpus linguistics*. (pp. 31-37). New York: Routledge.

Rhalmi, M. (2009). *Content-based instruction*. Accessed on 9th May 2015 at http://www.myenglishpages.com/blog/contentbased-instruction/

Richards, J. C. (2006). *Communicative language teaching today*. New York: Cambridge University Press.

Richards, J. C. & Rodgers, T. C. (2001). *Approaches and methods in language teaching*. New York: Cambridge University Press.

Rickheit, G., Strohner, H. & Vorwerg, C. (2008). The concept of communicative competence. In G. Rickheit and H. Strohner (Eds.) *Handbook of communication competence* (pp. 15-64). Berlin: Mouton de Gruyter.

Robinson, P. (1991). *ESP today: a practitioner's guide*. England: Prentice Hall International.

Römer, U. (2007). *Corpora and language teaching.* Retrieved August 25, 2015, from

http://www.uteroemer.com/HSK_Roemer_uncorrected_proofs.pdf

Rudzka-Ostin, B., Channel, J., Putseys, Y., & Ostin, P. (1985). *The words you need*. McMillan.

Salim, M.A.B., Ibrahim, N.A.B., Hassan, H. (2012). Language for tourism: A review of literature. *Procedia-Social and Behavioral Sciences*, 66, 136-143.

Salmani-Nodoushan, M. A. (2007). Are task type and text familiarity predicators of performance on tests of language for specific purposes? *The Asian ESP Journal*, 3 (1), 67-96.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access, *Neuropsychologia*, 20 (6), 615-627.

Schmitt N. (1997). Vocabulary learning strategies. In: N. Schmitt, M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy.* (pp. 199-227). Cambridge: Cambridge University Press.

Schmitt, N. & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, 19(1), 17-36.

Schmitt, N. (2000.) *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2005). Formulaic language: fixed and varied. *Elia,* 6 (6), 13-39.

Schmitt, N. (2008.) Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12, (3), 329-363.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26-43.

Schmitt, N. & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47, (4), 484-503.

Schmitt, N. & Zimmerman, C. B. (2012). Derivative forms: What do learners know? TESOL Quarterly, 36 (2), 145-171.

Scott, J. A. & Nagy, W. E. (2004). Developing word consciousness. In J. F. Baumann & E. J. Kame'enui, (Eds.), *Vocabulary instruction: Research to practice* (pp. 201-217). New York: Guilford.

Scowsill, D. (2017). *The economic impact of travel and tourism*. https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2017/world2017.pdf

Sharma, V. K., & Sharma, M. K. (2010). *Linguistic approaches to discourse analysis*. Lambert Academic Publishing.

Shin, D. & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English, *ELT Journal*, 64, 339-48.

Simion, O. M. (2012). The importance of teaching English in the field of tourism in universities, *Annals of the „Constantin Brâncusi" University of Târgu Jiu, Economy Series*, 2, 152-154.

Simpson-Vlach, R. & Ellis, C. N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31, (4), 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Developing linguistic corpora: A guide to good practice, Corpus and texts – basic principles.* Accessed 03/03/2019 at http://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The OSTI report*. UK: A & C Black.

Sinclair, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice,* (pp. 1-21). Oxford Text Archive.

Singer, M. (1990). *Psychology of language: An introduction to sentence and discourse processing.* Hillsdale, N J: Erlbaum.

Singh, L. K. (2008). *Fundamental of tourism and travel.* Gyan Publishing House.

Sketch Engine corpus software (Kilgarriff & Tugwell, 2001) available at

https://app.sketchengine.eu

Szirmai, M. (2001). *The theory and practice of corpus linguistics*. Debrecen: Debreceni Egyetem Kossuth Egyetemi Kiadója.

Szirmai, M. (2005). Bevezetés a korpusznyelvészetbe: A korpusznyelvészet alkalmazása az anyanyelv és az idegennyelv tanításában és tanulásában. Budapest: Tinta Könyvkiadó.

Sznajder, H. S. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes*, 29, 30-42.

Starfield, S. (2011). Doing critical ethnographic research into academic writing: The theory of methodology. In D. Belcher, A. M. Johns & B. Paltridge (Eds.), *New directions in English for specific purposes research*, (pp. 174-196). Ann Arbor: The University of Michigan Press.

Stevenson, A. & Waite, M. (2011). *Concise Oxford English Dictionary: Luxury edition*. Oxford: Oxford University Press.

Stoller, F. L. (2002). Promoting the acquisition of knowledge in a content-based course. In J. Crandall & D. Kaufman (Eds.), *Content-based instruction in higher education settings* (pp. 109–123). Alexandria, VA: TESOL.

Stoller, F. L. (2004). Content-based instruction: Perspectives on curriculum planning. *Annual Review of Applied Linguistic*s, 24 (1), 261-283.

Stoeckel, Ishii, & Bennett (2018). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*. https://academic.oup.com/applij/advance-article-abstract/doi/10.1093/applin/amy059/5253815?redirectedFrom=fulltext

Strevens, P. (1988). ESP after twenty years: A re-appraisal. In M. Tickoo (Ed.), *ESP: State of the art* (pp.1-13). SEAMEO Regional Language Centre.

Stubbs, M. (2001). *Words and Phrases*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology. In *English International Journal of Corpus Linguistics,* 7/2: 21, 5-244.

Sutarsyah, C., Nation, P. & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, 25 (2), 34-50.

Swales, J. M. (1971). *Writing scientific English*. Nelson.

Swales, J. M. (1985). *Episodes in ESP: A source and reference book on the development of English for science and technology.* Oxford: Pergamon.

Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

Swales, J. M. (1992). Language for specific purposes. In W. Bright (Ed.), *International encyclopaedia of linguistics* 2, pp. 300-302). New York, Oxford: Oxford University Press.

Takac, V. P. (2008). *Vocabulary learning strategies and foreign language acquisition.* Clevedon, UK: Multilingual Matters.

Tarnopolsky, O. (2012). *Constructivist blended learning approach to teaching English for specific purposes.* De Gruyter.

Terestyéni, E. (2009). Szaktudás idegen nyelven: A turisztika szakszókincsének rögzítési formái napjainkban. *Porta Lingua* 2009, pp. 395-402. Debrecen: Center Print.

Thornbury, S. (2002). *How to teach vocabulary.* London: Pearson Education Limited.

Todey, E. (2018). Introducing English for specific purposes. *TESL-EJ*, 22 (3), Available at http://www.tesl-ej.org/wordpress

Tono, Y. (2011). TaLC in action: recent innovations in corpus-based English language teaching in Japan. In A. Frankenberg-Garcia, G. Aston & L. Flowerdew (Eds.), *New trends in corpora and language learning*, pp. 3-25. London: Bloomsbury Publishing.

Tourism English Corpus. Available at https://bit.ly/3w1DWZC

Tribble, C. (1997). *Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching*. A paper presented at the First international conference: Practical Applications in Language Corpora (1997). University of Lodz, Poland. Retrieved August 28, 2010, from http://www.ctribble.co.uk/text/Palc.htm

Tutin, A. (2008). *For an extended definition of collocations*. Euralex: Barcelona.

Uribe, M. (2014). *Integrating the four English language skills through the CBI approach in a Colombian setting.* Presentation accessible at https://prezi.com/2thnmbkesjvu/cbi-skills-presentation/

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22 (2), 217-235.

Walker, R. & Harding. K. (2009). *Oxford English for careers: Tourism 3: Student's book*. Oxford: Oxford University Press.

Wang, F. (2010). *The four features of tourism English.* Retrieved 2nd November 2014 from http://www.seiofbluemountain.com/upload/product/201004/2010lyhy06a6.pdf

Wang, Z. (2014). A correlation analysis on the depth and breadth of ESL learners' vocabulary knowledge and their overall linguistic competence. *Theory and Practice in Language Studies*, 4 (12), 2460-2465.

Wanpen, S., Sonkoontod, K. & Nonkukhetkhong, K. (2013). Technical vocabulary proficiencies and vocabulary learning strategies of engineering students. *Procedia - Social and Behavioral Sciences,* 88, 312-320.

Webb, S. & Kagimoto, E. (2010). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics* December 17th, 1-19.

Weinert, R. (2011). Formulaicity and usage-based language: Linguistic, psycholinguistic and acquisitional manifestations. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication*. Chapter 1. Bloomsbury: London.

Wesche, M. B. (1993). Approaches to language study: Research issues and outcomes. In M. Krueger, & F. Ryan, (Eds.), *Language and content: discipline- content- approaches to language study*, pp. 80-95. Lexington, MA: D.C. Heath & Company.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

Widdowson, H. G. (1991). The description and prescription of language. In J. E. Alatis, (Ed), *Linguistics and language pedagogy: The state of the art*, pp. 11-24. Washington DC: Georgetown University Press.

Widdowson, H. (1998). Skills, abilities, and contexts of reality. *Annual Review of Applied Linguistics, 5 (18), 323-333.*

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21 (1), 3-25.

Wilkins, D. (1976). *Notional syllabuses: A taxonomy and its relevance to foreign language curriculum development.* London: Oxford University Press.

Williams, E. (1994). Remarks on lexical knowledge. *Lingua 92*, 7-34.

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151–171.

Williams, R. (1985). Teaching vocabulary recognition strategies in ESP reading. *English for Specific Purposes*, 4, 121-131.

Willis, D. (2003). *Rules, pattern, and words: Grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

World Tourism Organisation. (2014). *UNWTO tourism highlights*. https://www.e-unwto.org/doi/pdf/10.18111/9789284416226

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, 21 (4), 463-489.

Wray, A. (2002/a). Formulaic language in computer-supported communication: Theory meets reality. *Language Awareness*, 11 (2), 114-131.

Wray, A. (2002/b). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wouden, T. (2002). *Negative contexts: Collocation, polarity and multiple negation*. Routledge.

Zareva, A., Schwanenflugel, P. & Nikolova, I. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, 27 (4), 567-595.

Zipf, G. K. (1935). *The psychobiology of language: An introduction to dynamic philology.* Cambridge, Massachusetts: M.I.T. Press.

Zwiers, J. & Crawford, M. (2011). *Academic conversations: Classroom talk that fosters critical thinking and content understandings*. Stenhouse Publishers.

List of tables

List of figures

APPENDICES

APPENDIX A

*The first 100 words in the **Compleat Lexical Tutor Frequency Builder** output*

RANK	WORD	RANK	WORD	RANK	WORD
1.	THE	23.	YOUR	45.	PLACE
2.	AND	24.	VERY	46.	HAD
3.	A	25.	BE	47.	MY
4.	OF	26.	CAN	48.	OUR
5.	TO	27.	ALL	49.	MOST
6.	IS	28.	BUT	50.	WHICH
7.	IN	29.	BEACH	51.	HOTEL
8.	FOR	30.	AN	52.	FOOD
9.	YOU	31.	HAVE	53.	HAS
10.	WITH	32.	WILL	54.	BEACHES
11.	ARE	33.	BY	55.	SOME
12.	WAS	34.	ONE	56.	BEST
13.	IT	35.	GREAT	57.	ITS
14.	ON	36.	NOT	58.	THEY
15.	NUMBER	37.	OR	59.	GOOD
16.	THIS	38.	WERE	60.	EXPERIENCE
17.	AS	39.	CITY	61.	SERVICE
18.	THAT	40.	THERE	62.	JUST
19.	WE	41.	IF	63.	MORE
20.	I	42.	ALSO	64.	ROOM
21.	AT	43.	SPA	65.	TIME
22.	FROM	44.	SO	66.	STAFF

RANK	WORD	RANK	WORD
67.	**TOUR**	91.	**NICE**
68.	**OUT**	92.	**ROOMS**
69.	**MASSAGE**	93.	**WORLD**
70.	**BEAUTIFUL**	94.	**UP**
71.	**AREA**	95.	**REALLY**
72.	**LIKE**	96.	**TOURS**
73.	**WELL**	97.	**AMAZING**
74.	**GET**	98.	**DAY**
75.	**VISIT**	99.	**US**
76.	**MANY**	100.	**PEOPLE**
77.	**ENJOY**		
78.	**THEIR**		
79.	**HERE**		
80.	**FRIENDLY**		
81.	**PERFECT**		
82.	**RESTAURANT**		
83.	**WOULD**		
84.	**ABOUT**		
85.	**BACK**		
86.	**ONLY**		
87.	**WHEN**		
88.	**OTHER**		
89.	**WATER**		
90.	**ISLAND**		

APPENDIX B

Nouns (N) and verbs (V) of the full Tourism English Corpus wordlist

accommodation N	distance N	lounge N	staff N
airline N	diving N	luggage N	stay V/N
ambience N	eatery N	massage N	steam bath N
amenities N	enjoy V	monument N	serve V
apartment N	excess N	menu N	stroll V/N
appreciate V	excursion N	offer V/N	suggest V
archipelago N	expect V	order V/N	suite N
architecture N	experience N	palace N	sunbathing N
area N	explore V	passenger N	surround V
arrival N	facilities N	provide V	transfer V/N
atmosphere N	fare N	rafting N	transportation N
attraction N	fee N	rate N	toiletries N
baggage N	ferry N	recommend V	vibe N
beach N	flight N	region N	visit V/N
boast V	foodie N	relaxation N	villa N
book V	fountain N	rentals N	vineyard N
brasserie N	guest N	reputation N	voucher N
buffet N	guide N	reservation N	wave N
cancel V	hammam N	resort N	weather N
catering N	harbour N	return V/N	wildlife N
chef N	heritage N	robe N	winery N
climate N	honeymoon N	ruins N	wine tasting N
coast N	hospitality N	sailing N	worth V
compartment N	host N	salmon N	yacht N
complain V	include V	scenery N	
concierge N	indulge V	seafood N	
convenience N	interior N	season N	
cottage N	itinerary N	service N	
cruise N	journey N	shore N	
cuisine N	landscape N	sightseeing N	
delay V/N	lobby N	site N	
departure N	lobster N	skyline N	
destination N	location N	snorkelling N	
discount N	locker N	sommelier N	
discover V	lodging N	spa N	

APPENDIX C

Adjectives of the full Tourism English Corpus wordlist

accessible	hospitable	shallow
additional	ideal	situated
affordable	impressive	spacious
alluring	impeccable	spectacular
ancient	incredible	standard
appealing	international	stunning
aquatic	local	stylish
authentic	located	superb
available	luxurious	sublime
awesome	magnificent	traditional
breath-taking	major	tropical
bustling	marine	unforgettable
cosmopolitan	medieval	unique
cozy	memorable	unspoilt
culinary	mesmerizing	
delicious	natural	
delightful	nearby	
disappointed	picturesque	
domestic	pleasant	
excellent	pleased	
exclusive	popular	
exotic	pristine	
equipped	professional	
exquisite	reasonable	
fabulous	recreational	
famous	relaxing	
fascinating	renovated	
hidden	renowned	
historic	separate	

WORD LIST OF THE *AIR TRAVEL* SUBCORPUS

aboard	overhead locker
aircraft	passenger
airfare	passport
airline	rebook
airport	reclaim
aisle	refund
allowance	reimburse
altitude	restrictions
amenities	route
aviation	runway
baggage	safety instructions
boarding	seatbelt
boarding pass	service
cabin crew	shuttle bus
carry-on	skybus
check-in	staff
compartment	suitcase
connection	terminal
conveyor belt	transfer
economy class	transport
emergency	turbulence
exceed	usher
flight	via
flight attendant	visa
gate	voucher
jet lag	zone
limit	
lounge	
luggage	
name tag	

*WORD LIST OF THE **BEACH HOLIDAYS** SUBCORPUS*

abroad	explore	pristine
affordable	facilities	reef
ambience	fauna	seashore
amenities	ferry	snorkelling
anchor	flora	stroll
angling	getaway	sunbathe
aquatic	gulf	sunscreen
archipelago	hammock	tan
bathe	harbour	wave
beachgoer	holidaymaker	yacht
breeze	honeymoon	
canoeing	hospitality	
Caribbean	houseboat	
climate	hurricane	
convenience	indulge	
coral	isle	
cottage	isolated	
crab	itinerary	
crew	lagoon	
cruise	landmark	
crystal	landscape	
cuisine	lifeguard	
culinary	lighthouse	
currents	lobster	
deck	mainland	
depth	marine	
dive	mesmerizing	
dockyards	mild	
dolphin	parasailing	
drift	peninsula	
ecosystem	pier	
embark	port	

APPENDIX F

*WORD LIST OF THE **CITY TOURS** SUBCORPUS*

architecture

boast

cathedral

classical

cobbled

cosmopolitan

court

crowded

cuisine

culinary

depository

display

district

downtown

dungeon

eclectic

enchanting

equestrian

escort

ethnic

exhibition

facade

facilities

fare

fortress

fountain

fresco

funicular

gothic

guide

historical

ornament

palace

parliament

pavilion

pedestrian

piazza

pickpocket

picturesque

pigeon

plaque

priceless

Romanesque

route

ruins

sanctuary

scenery

sculpture

sidewalk

sightseeing

skyline

skyscraper

souk

souvenir

spectacular

stadium

statue

streetcar

stroll

stunning

style

subway

tavern

temple

tomb

traffic

tram

underground

uptown

vibrant

vista

workshop

*WORD LIST OF THE **HOTELS AND ACCOMMODATION** SUBCORPUS*

accessible	duvet	overlooking
accommodate	eatery	overrated
affordable	elevator	parking lot
ambience	equipped	pillow
amenities	facilities	pleased
apartment	faucet	plug
architecture	function room	pricey
arrival	furnishing	room rate
atmosphere	guest	reduction
attentive	hallway	refund
available	hospitality	registration
baggage	host	rentals
balcony	hostel	reservation
ballroom	housekeeper	sheet
banquet	impeccable	socket
bartender	include	spacious
bathrobe	inn	stay
bedding	interior	studio
belongings	kitchenette	stylish
closet	lavatory	suitcase
concierge	laundry	suite
cozy	lobby	toiletries
cottage	locker	valuables
curtains	lodging	vibe
decor	lounge	vicinity
disappointed	luggage	wardrobe
discount	luxurious	
disturb	maid	
dorm	maintenance	
dresser	mattress	

*WORD LIST OF THE **RESTAURANTS AND CATERING** SUBCORPUS*

ambience	hospitality	spirits
appetiser	host/hostess	squeezed
atmosphere	impeccable	staff
attentive	leftover	starter
banquet	lobster	steamed
bartender	marinated	stew
beverages	mouth-watering	stuffed
bill	mutton	sundae
braised	napkin	tablecloth
brasserie	nourishing	tasteful/tasteless
brew	offer	tavern
brunch	order	tender
buffet	organic	treat
champagne	oyster	veggies
chilled	pastries	venison
complain	platter	vibe
course	poached	vinegar
crab	portion	
crispy	pour	
cuisine	prawn	
culinary	properly	
cutlery	ratatouille	
delicacies	recommend	
delicious	reservation	
dine	roast	
dish	salmon	
flavourful	shrimp	
fried	sommelier	
gourmet	soufflé	
gravy	speciality	

*WORD LIST OF THE **WELLNESS AND SPA** SUBCORPUS*

acupressure	indulge
affordable	lifeguard
ambience	locker
amenities	masseur/masseuse
aqua gym	medical
aromatherapy	mud bath
atmosphere	nude
attentive	overrated
balneotherapy	pamper
bathe	physiotherapy
bathrobe	reflexology
belongings	refresh
blissful	refund
bubble bath	regeneration
change room	relieved
chlorinated	resort
client	sanitation
decor	sauna
degree	scent
delightful	sunbed
discount	swimwear
dissatisfied	tanned
equipped	therapist
facilities	toiletries
hammam	treatment
healing	unwind
herbal	vibe
hydro massage	voucher
hygiene	wellbeing
impeccable	whirlpool

LIST OF THE TOP FORTY MOST FREQUENT THREE- AND FOUR-WORD BUNDLES IN THE TOURISM ENGLISH CORPUS

RANK	Three-word	Frequency	Four-word	Frequency
1	one of the	267	is one of the	73
2	some of the	146	one of the most	61
3	if you are	113	one of the best	58
4	of the most	108	some of the most	41
5	as well as	95	in the heart of	35
6	of the best	86	if you want to	34
7	is one of	82	all over the world	31
8	of the city	72	from all over the	26
9	in the world	62	some of the best	21
10	it is a	61	as one of the	20
11	the food was	57	you are looking for	19
12	a lot of	50	a great place to	18
13	the heart of	49	is located in the	18
14	you want to	49	are some of the	17
15	the food is	47	at the end of	17
16	the city is	44	if you are looking	17
17	there is a	44	of the most popular	16
18	this is a	44	the heart of the	16
19	view of the	44	will definitely be back	16
20	and it was	43	as well as a	15
21	if you want	43	in the middle of	15
22	the service was	41	of the city's	15
23	you can also	39	when it comes to	15
24	all over the	38	a few of the	14
25	located in the	38	one of the top	14
26	this is the	38	the best way to	13
27	you can enjoy	37	will be able to	13
28	you will be	37	a wide range of	12
29	a variety of	36	as well as the	12
30	of the world	36	in the United States	12
31	on the beach	36	of the most famous	12
32	in the heart	35	the perfect place to	12
33	part of the	34	there are plenty of	12
34	most of the	33	a view of the	11
35	the hotel is	33	can't wait to	11
36	very friendly and	32	from start to finish	11
37	in the area	31	I've ever had	11
38	in the city	31	in the center of	11
39	is located in	31	is a great place	11
40	over the world	31	of some of the	11

LIST OF THE TOP 50 MOST FREQUENT TOURISM-SPECIFIC BUNDLES IN THE TOURISM ENGLISH CORPUS

RANK	TOURISM-SPECIFIC BUNDLES	RANK	TOURISM-SPECIFIC BUNDLES
1	beach resort located on	26	to make your stay
2	the food was delicious	27	white sand beaches and
3	views of the city	28	with information on hotels
4	you can enjoy the	29	would definitely come back
5	indoor and outdoor pools	30	a perfect tropical paradise
6	one of the best beaches	31	a popular tourist destination
7	a jet boat ride	32	and thrilling water sports
8	and the service was	33	and watch the cruise
9	day tours of Rome	34	beaches offering family fun
10	tourists from all over	35	breath-taking view of the
11	with a view of	36	for travellers like you
12	guided tour in Madrid	37	guests can enjoy a
13	a perfect place for	38	highly recommend this restaurant
14	inn suites in Miami	39	included in the price
15	never fails to lure	40	is a hidden gem
16	a delightful and memorable	41	is an online travel
17	and explore some of	42	majestic palm trees gently
18	city tours offer a	43	most of the tourists
19	destinations in the world	44	of tourists and vacationers
20	drinks and snacks for	45	offers food and drink
21	just want to relax	46	paradise filled with deliciousness
22	of the amenities offered	47	pool and hot tub
23	staff are very friendly	48	quality of the food
24	the best foodie vacations	49	relax and enjoy the
25	the most visited cities	50	rentals and shower facilities

APPENDIX L

THE MOST FREQUENT LEFT AND RIGHT BINARY AND FOUR-WORD SPAN COLLOCATES OF NOUNS IN THE TOURISM ENGLISH CORPUS WORDLIST

Rank	Left collocate	NODE	Right collocate	Rank	Left collocate	NODE	Right collocate
1	local, comfortable, budget, excellent, fine cheap, affordable, save on, selection of, range of	ACCOMMODATION	facilities, rate, to vacationers, at affordable price, to global tourists	58	swim, diving	SNORKELLING	great
2	connecting, cancellations for, itinerary for, schedule of the	AIRLINE	baggage rules, 's weight allowance, ticket websites	59	excellent, attentive, staff, cocktail, menu	SOMMELIER	know, provide, serve
3	wonderful, welcoming, cozy, relaxing	AMBIENCE	with attentive staff, was satisfactory	60	de-stress, perfect, outdoor, relax, herbal, thermal, beer	SPA	package, facilities, sauna massage, hammam, steam
4	word-class, upscale, ultra-modern,	AMENITIES	available, include, offer	61	excellent, helpful, attentive front desk, friendly, maid	STAFF	knowledgable, polite treat helpful, careful, friendly, professional, courteous
5	private, contemporary,	APARTMENT	hotel, is equipped	62	definitely, place to, enjoy	STAY	overnight, apartment, room
6	islands in the, delightful	ARCHIPELAGO	extends for, divided by	63	sauna	STEAM BATH	fitness room
7	spectacular, heritage, Romanesque and Gothic,	ARCHITECTURE	is beguiling	64	leisurely, relaxing,	STROLL	tropical paradise, vineyard, mosque
8	seating, wellness, coastal, dining, metropolitan	AREA	overlooking the ocean, which is accessible	65	scuba, fishing, shark, snorkelling	DIVING	boat tour, lessons, beach experience, parasailing
9	welcome on, upon,scheduled	ARRIVAL	and departure, process	66	spacious, stylish,favourite	EATERY	classic,seats,offers, open
10	décor and, festive and cozy, relaxing, pleasant, calm	ATMOSPHERE	with live music, warm, charming, elegant	67	check-in, luggage, charge	EXCESS	pieces, luggage, baggage, fee
11	enchanting, tourist, major	ATTRACTION	to tourists/visitors	68	city,nature,shore,fabulous	EXCURSION	unique, tour, explore
12	carry-on, mishandled, checked, lost	BAGGAGE	fee, allowance, handling storage, tag, policies	69	spa, dining, relaxing, outdoor, nightlife,culinary	EXPERIENCE	provide, memorable, local, traditional, perfect
13	accessible, hidden, fabulous, luxury, white sandy, surf,	BEACH	that can captivate, with soothing/crystal water	70	spa, luxurious, additional, conference, basic, rental	FACILITIES	equipped, amenities, include
14	traditional, French	BRASSERIE	and wine bar, steakhouse	71	standard, one-way, shuttle, diner, local	FARE	service
15	breakfast, free	BUFFET	with separate stations,	72	additional, charge, baggage, entrance	FEE	necessary
16	self, railway	CATERING	needs	73	take, go by	FERRY	terminal, service, system
17	expert, celebrity, four-star, head	CHEF	restaurateur, 's special	74	book, domestic, available, connecting, round-trip	FLIGHT	time, attendant, delay, cancellation, connection
18	tropical, hospitable, warm, pleasant, mild, balmy, temperate	CLIMATE	weather	75	ice	FOUNTAIN	areas
19	along the, off the, west, east, south	COAST	line, region, resort	76	individual, hotel	GUEST	service, room, enjoy

20	luggage, passenger, first-class	COMPARTMENT	for personal items	77	tour, travel, professional, local, expert	GUIDE	commentary, lead, recommend, take
21	hotel features a	CONCIERGE	service, levels	78	relaxing, sauna, jacuzzi	HAMMAM	bath, treatment,experience rejuvenation,pool,sauna
22	comfort and, added,high-tech	CONVENIENCE	store	79	fishing, around the	HARBOUR	area, bridge, trip
23	vacation	COTTAGE	at affordable price	80	cultural, historic, world commercial, musical	HERITAGE	site, village, renowned, monuments, architecture
24	sightseeing, boat, river, sunset, houseboat, barge	CRUISE	holiday, ship, itinerary, ride, travel, provide	81	newlywed, attract, suites, specialized for, ideal	HONEYMOON	packages, couples, resort, destination, vacation, trip
25	homeland, local, sophisticated authentic, word-class, superb, exquisite,continental, excellent	CUISINE	bar, restaurant, present, include	82	kindness, warm, comfort, full-service, traditional, unparalleled	HOSPITALITY	experience,industry,given, could be improved
26	ask about	DELAY	before	83	offers, superb	HOST	of activities, of cultural attractions, next meeting
27	flight's, scheduled, arrival, original	DEPARTURE	point, station, online	84	create, soft, modern, glistening, contemporary	INTERIOR	design, casual, quaint
28	tourist, final, all-year- round, popular, holiday, top, amazing ideal, perfect, wonderful, next	DESTINATION	airport, chosen, known, offers	85	popular, food, price	ITINERARY	includes, change, confirmation
29	special, great, percent, try	DISCOUNT	airline, carrier	86	embark on, worth, return, spa	JOURNEY	to, across, through
30	walking, reasonable, long, short, driving	DISTANCE	travel, between, from	87	rural, colourful, volcanic scenic, marvellous	LANDSCAPE	sceneries, dominating, in, of
31	body,treatment,relaxation,deep hot stone, spa, muscle, aroma	MASSAGE	experience, followed by session,requested,steam	88	unique, ultra-modern, hotel	LOBBY	area, restaurant
32	marvellous	MONUMENT	referred to	89	freshly caught, seafood tower with, exceptional	LOBSTER	crab, meat, avocado,sauce
33	seasonal, dessert, excellent, multi-course, a la carte, tasting	MENU	includes, involving, options, item	90	preferred, perfect, scenic convenient, unique,ideal	LOCATION	close to, downtown, between
34	holiday, room, excursion, supposed to, island, resort, hotel,	OFFER	package, selection, choice, service, variety, getaway	91	into the, shower	LOCKER	area, room, facilities
35	preference, treatment, working, place	ORDER	dish, drink, online	92	lobby's, bar, restaurant, relaxation, comfy	LOUNGE	style, with infrared light
36	royal,17th-century, tour of the, Turkish style	PALACE	stunning, super-luxurious, definitely	93	family, master, king, presidential, luxurious	SUITE	including, standard, with a balcony
37	permitted per, procedure for, screen, baggage	PASSENGER	check, carrier, comfort, compartment,load,plane	94	check, pick up, carry-on, lost, piece of, excess, deliver, weight of,packing	LUGGAGE	allowances, compartment, exceeds, packing, policies, tag, tip, transfer
38	white-water, fishing, diving	RAFTING	down the rapids, deep-sea fishing	95	treat for, do, sightseeing and, activities such as	SUNBATHING	relaxing, swimming, watersports
	room, economical, competitive flat, high, first,	RATE	for, at hotels	96	group, fast	TRANSFER	to, of
40	spectacular, enchanting, surrounding the, explore	REGION	attracts, comprising	97	public, air, means of	TRANSPORT	system, from

#				#				
41	perfect for, complete, provide, spa, wellbeing, massage	RELAXATION	area, experience, massage, room, lounge	98	bathroom with, without minimal, stop stealing, eco-chic	TOILETRIES	from, as	
42	vacation, sport, yacht, beach	RENTALS	water sports, include	99	upbeat, relaxed, casual, friendly, cool, unpretentious	VIBE	décor, guest, music, seating	
43	world-wide, built up its	REPUTATION	for, and reference	100	well worth, decided to, would definitely, enjoy, never miss to, tourists, best time to, plan to	VISIT	again, explore, during	
44	make, need, online, without	RESERVATION	for, needed	101	rent, stay, beautiful, holiday	VILLA	on, with	
45	popular, beach, welcoming, coastal, beachside, all inclusive	RESORT	offers, located	102	sight of the, city	VINEYARD	offer, bring	
46	given a, cosy, dressing	ROBE	slippers, towel, change	103	gift, food, entry, appropriate, possible	VOUCHER	for, amounts	
47	ancient, extensive, amazing, Roman, visit includes	RUINS	of the, which are situated	104	hollowest	WAVE	surfing, chasing, carving that crashes the shores	
48	boat tours, sunset, traditional include, para-,	SAILING	boat, excursion, holiday, houseboat, canoeing	105	sunny, Mediterranean, comfortable, tropical, cold bad, rainy, warm, cloudy	WEATHER	during, in, of, with, was	
49	grilled, smoked, perfect, Irish	SALMON	caviar, spicy, pasta,	106	incredible, varied, admire abundant, national, exotic exquisite, indigenous	WILDLIFE	reserves, sanctuaries, park refugee, tours	
50	unspoiled, picturesque, jaw-dropping, spectacular, stunning dramatic, fascinating, fantastic	SCENERY	combining, on, of	107	small-group	WINERY	tour	
51	indigenous, fresh, menu of, array of, flavorful, include	SEAFOOD	mindblowing, pasta, restaurant, guaranteed	108	includes	WINE-TASTING	lunch	
52	peak, holiday, warm, travel, holiday, high, tourist, vacation	SEASON	means, when	109	hire, rent, take, chartered	YACHT	club, rentals, scuba diving	
53	excellent, great, enjoyed, room food, maid, customer, dedicated superb, friendly, personalized, fabulous, courteous, no-frill	SERVICE	ambience, atmosphere, after, beverages, provided, always, team delicious, attentive, counter, expected					
54	sandy, sea, off-, visit, tour	SHORE	beaches, excursion, patrol					
55	memorable, narrated, extensive incredible, intriguing, quality, informative, first-class, the best	SIGHTSEEING	adventure, cruise, option company, destination, itinerary, package, tour					
56	heritage, popular, located, web	SITE	activities, based					
57	city, constructing	SKYLINE	boasts					

The most frequent left and right binary and four-word span collocates of verbs in the Tourism

English Corpus wordlist

Rank	Left collocate	NODE	Right collocate
1	can/cannot, will, don't	APPRECIATE	ocean breeze, proximity, beauty, amenities, luxury
2	properties, suites, can	BOAST	with flat screen TV, view, fully furnished kitchen
3	must, used to, need to, should, advisable to, make sure to	BOOK	in advance, online, hotel, ticket, vacation, accommodation, flight, tour
4	option to	CANCEL	flight, bookings
5	passengers, fliers	COMPLAIN	about
6	ask about the	DELAY	before
7	to, be able to, will, long to	DISCOVER	ancient city, beautiful sight, best food, vibrant culture, aquatic paradise, uninhabited island, fascinating city
8	can, will, would, really, fully, surely, visitors, relax and	ENJOY	meal, dining experience, beach, view, attractions, facilities, pool, diversity, nature, hammam, scuba diving
9	can, could, would, do not, might, certainly	EXPECT	to pay, to find, lot to explore, from a five-star,
10	must, can, will, to	EXPLORE	hidden beauty, coastline, region, flora and fauna, coral reefs, heritage
11	amenities, tours, packages, bathrooms, services, facilities, accommodations	INCLUDE	travel, direct flights, air conditioning, swimming pool, Internet connection
12	chance to, time to, lot more to	INDULGE	food tasting, culinary experience, yourself
13	hotels, rooms, tours, resorts, beaches, airlines, excursions, packages, cruise, merely	OFFER	dream/magic vacation, opportunities, incredible rooms
14	take, place, decided to, must	ORDER	dish, wine, juices, mussels
15	resorts, vacation, landscaping cottages, island,	PROVIDE	relaxation, service, accommodation, spaces rooms, perfect setting, accommodation, information, view
16	highly, definitely, totally, surely, absolutely, really, thoroughly	RECOMMEND	service, place, restaurant, spa, treatment, travel, experience
17	definitely, love to,	RETURN	to, home
18	would, must, should, place to,	STAY	vacation/private apartment, hotel, motel, overnight,
19	resorts, operators,	SERVE	traditional dishes, outstanding breakfast, portions
20	leisurely, relaxing	STROLL	along, through, around, down
21	definitely, strongly, highly	SUGGEST	that you, the very best
22	churches, architecture	SURROUND	square
23	definitely, frequently, want to, must, never miss to, destinations	VISIT	place, island, attraction, beach, temple
24	well, exactly, definitely, truly,	WORTH	exploring, visiting, seeing, price, journey

The most frequent left and right binary and four-word span collocates of adjectives in the Tourism English Corpus wordlist

Rank	Left collocate	NODE	Right collocate	Rank	Left collocate	NODE	Right collocate
1	easily, fully	ACCESSIBLE	location, on foot, by boat, by road	38	comfortably, wonderful, spacious, super, very	LUXURIOUS	spa, seats, villas, facilities holiday, hotel, beach resort
2	avoid, aware of	ADDITIONAL	fee, services, facilities nights, connections	39	it is a	MAGNIFICIENT	building, view
3	offers, relatively, more	AFFORDABLE	accommodation, price, deal, rates	40	one of the, most of its, some of the, several, all the	MAJOR	destination, attraction, airline, sight, event, region, operator, city
4	other, its	ALLURING	vacation spots, beaches, island, reefs	41	exotic, abundant, brilliant exciting, variety of	MARINE	life, conservation area, preserve, sport, animals
5	discover, combination of, blend of, hidden	ANCIENT	civilization, culture, ruins, churches, walls, buildings	42	ancient, charming, famous, preserved,	MEDIEVAL	town, castle, cathedral, village, origin
6	offered by, more, most	APPEALING	resorts, attractions, islands	43	truly, most, more	MEMORABLE	experience, vacation, activities, trip, gift
7	various, other	AQUATIC	diversity, flora and fauna, paradise, sports	44	indeed	MESMERIZING	beach, holiday, place, vacation, coastal village
8	experience, discover, try, taste	AUTHENTIC	food, taste, cuisine, massage, places	45	wonderful, extreme, spellbinding, outstanding incomparable, fantastic	NATURAL	attraction, beauty, hot springs, ingredients, surroundings, sites
9	equipment, parking, facilities, lockers, beverages	AVAILABLE	flight, option, to guests vegetarians, visitors	46	gift shops, restaurants, beaches	NEARBY	accommodation, camp, beach, restaurant, tour
10	experience, food, service, really	AWESOME	time, place, stay, spa, selection, service, staff	47	known for, most,	PICTURESQUE	backwaters, bayside, town, beach, scenery, square
1	enjoy, visit, view, beach, charm of	BREATHTAKING	sight, view, beach, landscape, architecture	48	extremely, very, more, most	PLEASANT	experience, vacation, weather, room, walk
12	vibrant, place, square	BUSTLING	market, restaurant, neighbourhood, area	49	extremely, very	PLEASED	to, with, choice, treatment, room, pool
13	sparkling, hotel, modern, sophisticated	COSMOPOLITAN	city, culture, destination, traveller	50	getaway, very, extremely most, more, increasingly	POPULAR	activities, historic sites, city, attraction, bar, beach
14	ambience, area, environment, room, warm, very	COZY	restaurant, room, place, table, atmosphere, interior, ambience	51	relax, visit, experience	PRISTINE	beach, lake, waterfall, rushing river, sand
15	enjoyable, exciting, elevated, exceptional, excellent, distinguished, bonafide, special, exquisite	CULINARY	experience, destination art, expert, delight, creations	52	very, totally, utterly, skilled, highly, extremely, exceptional,	PROFESSIONAL	attentive, courteous, conscientious, friendly, knowledgable, polite, welcoming, staff, guide
16	unbelievably, particularly, inventive, absolutely, tasted, completely, really, fresh-made	DELICIOUS	food, sauces, cocktail, dinner, breakfast, bread, flavours, home-made vegetarian, meal, dessert	53	fairly, quite, very, beyond, highly, offered at	REASONABLE	price, fee, tariff rates expectation, distance,

#				#			
17	offers, absolutely	DELIGHTFUL	historical town, sights, shoreline, ambience, spa treatment, place, island	54	promotes, good	RECREATIONAL	activities, areas, beach, trails
18	really, a bit, a little, very	DISAPPOINTED	because, but, with	55	atmosphere, wonderfully truly, cosy, view, enjoy	RELAXING	ambience, enjoyable, quiet massage, atmosphere, day experience, holiday, spa
19	international and, regions of growth in, lost on a	DOMESTIC	destination, flight, tour travel, passengers	56	newly, completely, fully building	RENOVATED	well-maintained, museum, room, hotel, motel
20	provide, service, tasted, food, experience, restaurant, offer	EXCELLENT	accommodation, staff, service, resort, location, facilities, hospitality, food, massage, meal,spa	57	internationally, word-, including, island	RENOWNED	destination, thermal spa, trekking spot, beach resort chef-restaurateur, architectural works
21	receive, offer,	EXCLUSIVE	atmosphere, access holiday, service, treat	58	spa, with, villas feature, seating areas, buffet with	SEPARATE	cottages, houses, stations, premise for luggage, area,
22	picturesque, famous,	EXOTIC	destination, beach, bird beauty, food, island, flavour, resort, wildlife	59	relatively	SHALLOW	ocean, pool, river, waters tourist city
23	completely, well, fully	EQUIPPED	bar, kitchen, with	60	centrally, discreetly,	SITUATED	close to, near, within
24	food, offer, absolutely	EXQUISITE	cuisine, dining, food, selection, wildlife	61	quite, relatively, very, provide	SPACIOUS	room, suite, terrace
25	absolutely, pool, dessert, dish super, enjoy, drink choice	FABULOUS	beach destination, city attraction, experience, food, staff, holiday	62	environment, service, admire, known for, historical	SPECTACULAR	sights, architecture, city beach resort, coastal, sight landscape, landmark
26	most, very, word, globally	FAMOUS	island, cathedral, castle monument, city, view beach resort, temple	63	high-class, suite	STANDARD	service, room, timing
27	enjoy	FASCINATING	architecture, history, beach resort, city, sight, scenery,sunset,sculpture	64	palace, coast, area, décor, place, sight, offer, enjoy, truly, absolutely	STUNNING	attraction, beach,cityscape vista, collection,coastal walk, scenery,view,venue
28	experience, explore, perfect recommend, spa	HIDDEN	away,beach,beauty,gem corner,coves,treasure	65	décor, genuinely, serving	STYLISH	décor, dishes, eatery, room
29	superb, impressive, most, explore, rich	HISTORIC	ambience, building, site centre, church, heritage, landmark, monument	66	offers, service, facilities, food, menu	SUPERB	accommodation, cuisine, dining choices, service, experience, massage
30	exceptionally, especially,super	HOSPITABLE	staff, climate	67	food, dish, volcano	SUBLIME	festival, events, wine
31	find, beach, resort, spa, offer	IDEAL	choice,destination,place honeymoon, location,	68	served in, flavoured of, enjoy, serve, selection of	TRADITIONAL	environment, architecture, dish, cuisine, massage
32	range of	IMPRESSIVE	authentic, historic,menu sight, view, quantity	69	exotic, surrounded by, perfect, unspoiled	TROPICAL	climate, ambience,weather hideaway, paradise, island
33	cuisine, service, staff, interior	IMPECCABLE	service, staff	70	vacation tour, visit, create perfect, honeymoon	UNFORGETTABLE	experience, stay, sunset, day, vacation, honeymoon
34	food, absolutely, offer, course, explore, find	INCREDIBLE	affordable, service, spot attention,buildings,view	71	provide, enjoy, amazing, experience, offers, truly	UNIQUE	attraction, destination, spa, adventure, cruise, charm
35	tasty, several, lack of, exciting	INTERNATIONAL	airport, cuisine, travel, destination, flight,spot	72	perfect blend of, consists of	UNSPOILT	beach, island, location
36	lovely, fresh, enjoy, variety of, authentic, experience,	LOCAL	accommodation, craft apartments, atmosphere				
37	centrally,excellently,island,spa hotel,resort, restaurant,beach	LOCATED	directly, near, right, along				

Task type A:

Main objective: to enhance learners' collocational awareness of key tourism lexical items

A1: Study the following collocational profile of the verb INCLUDE and underline the nouns that are likely to be used with this verb in tourism context

amenities, organizations, directives, tours, packages, meanings, bathrooms, requirements, characters, actions, services, facilities, membership, accommodations, direct flights, insurance, benefits, air conditioning, topics, swimming pool, instructions, Internet connection

A2: Add five more nouns that are often used with the verb INCLUDE in tourism language

a.

b.

c.

d.

e.

Key to Task type A1:

amenities, tours, packages, bathrooms, services, facilities, accommodations, direct flights, insurance, air conditioning, swimming pool, Internet connection

Key to Task type A2:

  a.   breakfast b. ticket c. transportation d. en-suite bathroom e. minibar

Task type B:

Aim: to help learners observe and identify collocates of lexical items in tourism context

B1: After studying the concordance lines fill in the sentences below with the appropriate words

is quaint and has an intimate ambience. Food and wine pairing was impeccable. A return to this excellent restaurant was well over due s
tdoor pool is cooler and my two teens loved both. The grounds are impeccable and there are roses and landscaping, outdoor picnic area
?must?. We come to Spa once a year just for the thermes. Service is impeccable especially if you book one of their packages. We usually
ing on in there. A perfect ambiance for us. The service was casually impeccable; our server was friendly, spoke some Italian, and made u
scayne. With panoramic views of Biscayne Bay, world-class cuisine, impeccable service and laid-back luxury at every turn, Hotel Aria, BW
ter than what we were anticipating. The place is vibrant, dim lit, and impeccable service is at the heart of this restaurant. My wife and I we
m of an Italian restaurant. Authentic local spot with great food and impeccable service. Recommendations spot on and our waiter was f
us, the atmosphere was romantic and peaceful, and the service was impeccable! South Africa?a place like no other on the planet. It?s the
d level of conversations throughout the restaurant. Greetings by an impeccable staff preceded the introduction to this "Hollywood" set. (
t to finish, the whole experience here was amazing. The service was impeccable, the food was divine (we opted for the 4 course menu) a
d is peaceful and quiet, even with lots of people around. The staff is impeccable. The location of the hotel is great with downtown Willem
lid not disappoint.From start to finish, the staff and the dishes were impeccable. The very first canap\x82 evoked the seaside growing up
onverted from a 19th-century sandstone mansion. The interiors are impeccable, the view from every room is magnificent, the staff are p
ry the butter cake - it absolutely melts in your mouth. The service is impeccable. This place is a bit pricey, but worth every penny for a sp
e was paired with wine. Everything was served family style. Service- Impeccable! We had the same server as we did during our last visit (
ecial to celebrate, try Bosk. The decor is stunning, and the service is impeccable. We ordered the four-course meal which had the most u
be back, everyone was so friendly and welcoming and service was impeccable. We were so blessed to have made this selection, as our

a, The décor is stunning, and the _____ is **impeccable**.

b, Food and _____ was **impeccable**.

c, The _____ is **impeccable,** the view from every room is magnificient.

d, Greetings by an **impeccable** _____ preceeded the introduction of this 'Hollywood' set.

e, Service is **impeccable,** _____ if you book one of their packages.

Key to Task type B1:

a, The décor is stunning, and the *service* is **impeccable**.

b, Food and *wine pairing* was **impeccable**.

c, The *interiors* are **impeccable,** the view from every room is magnificient.

d, Greetings by an **impeccable** *staff* **preceeded** the introduction of this 'Hollywood' set.

e, Service is **impeccable,** *especially* if you book one of their packages.

Task type B2: Observe the concordance lines and list verbs that are used with the noun LUGGAGE

art of airline and airport personnel responsible for keeping track of luggage. According to the U.S. Department of Transportation, U.S. a
ago - Sacramento, for example, you'll need to put it in your checked luggage after customs in Chicago?. 2. The gate change effect Check
similar policies but low cost airlines run the gamut from generous luggage allowances to as little as 30lbs per passenger. You may not
/interlining agreements which means you will have to pick up your luggage and check in with the next airline. This is especially prevaler
rtran flight was too, but she was not aware of having to pick up her luggage and check it again, and to check herself in for the flight. Mal
nger check before entering the gate, boarding, flying and pick-up of luggage and - limited to international flights - another border contro
onsible for the increasing amount of misplaced and downright lost luggage around the globe. Air Jamaica lost my checked bag on a flig
ne may simply not be able to accommodate the same size carry-on luggage as a jet. Check with the individual airline you are flying rega
w to Chicago, don't be surprised if British Airways only checks your luggage as far as London. The responsibility for the luggage primari
If you need assistance to get to your gate check in as much of your luggage as possible. It not only makes it faster to help get you there
. The pick-up line - for luggage When you are first checking in your luggage, ask if you need to pick up luggage - ie. clearing customs m
t you may want to consider. From making connections to will your luggage be checked through - connections are often a beast!? The ru
ith. Air Canada and Delta are not partner airlines but you can check luggage. Beach holidays corpus Holidays on Paradise beach on the
ra pound or kilogram.   If you travel on more than one airline your luggage can be booked through- Outside of the weight of luggage,
the passenger agent know that you are connecting, and ask if your luggage can be checked through to your destination.   If you are wi
a great way to have peace of mind on your flight. The size or your luggage can lead to fees and problems if it is not the right size. Ame
ht restriction, which means that the total weight of your checked in luggage can not exceed 20kg, or 44lbs.   On a paper ticket, you can
cy airline like United Airlines to a low cost one like Aer Lingus your luggage cannot be checked through, and neither can you so you nee
ou. You can reclaim it either at the gate or with the regular checked luggage. Carefully edit your carry-on baggage Pack valuable items
y. If you are flying domestically with luggage that has to go into the luggage compartment, expect to pay baggage fees on all but a coup

Key to task type B2: keep track of, pick up, check in, reclaim, exceed, put in

Task type C:

Aim: practice the collocational patterns of particular tourism keywords

Task: Complete the sentences with the missing words. The same word should be used in one set of concordance lines.

C1. *missing word*: _____

styles and lavish landscape provide a truly breathtaking _____ of Miami's Latin Quarter

The volcano is an amazing _____ in itself, towering to 12,000 feet high.

This tour will give you an overall impression of the most significant _____of Vienna.

cruise by a traditional boat and enjoying the _____of a vineyard-lined countryside

lounging around and just taking pleasure in the beautiful _____ while having a snack

C2: *missing word*: _____

The castle itself is huge and needs several hours to _____ properly.

Come and _____ some of the most magnificient piazzas, ancient buildings and

Glass bottom boats let you _____ the bottom of the beautiful ocean floor.

We took a full morning to _____ and still didn't see all the museums, treasures and

Be it the beaches or the hinterland the best way to _____ all the attractions is car rental.

C3: *missing word*: _____

Our meals were _____ and all three of us were entirely satisfied with the food.

clean spa with friendly service, the massage _____ and fully explained before the start

Today the lake offers an _____ boating spot for the tourists.

The wonderful Tuscany, Italy, is an _____ choice of holiday destination.

praised as the best hotel in Budapest and has _____ reviews on booking websites

Key to task type C: 1. sight 2. explore 3. excellent